cemmap
centre for microdata methods and practice

Adapting to misspecification

Timothy B. Armstrong
Patrick Kline
Liyang Sun

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP18/24

Economic and Social Research Council

Adapting to Misspecification*

Timothy B. Armstrong† Patrick Kline‡ and Liyang Sun§

August 2024

Abstract

Empirical research typically involves a robustness-efficiency tradeoff. A researcher seeking to estimate a scalar parameter can invoke strong assumptions to motivate a restricted estimator that is precise but may be heavily biased, or they can relax some of these assumptions to motivate a more robust, but variable, unrestricted estimator. When a bound on the bias of the restricted estimator is available, it is optimal to shrink the unrestricted estimator towards the restricted estimator. For settings where a bound on the bias of the restricted estimator is unknown, we propose adaptive estimators that minimize the percentage increase in worst case risk relative to an oracle that knows the bound. We show that adaptive estimators solve a weighted convex minimax problem and provide lookup tables facilitating their rapid computation. Revisiting some well known empirical studies where questions of model specification arise, we examine the advantages of adapting to—rather than testing for—misspecification.

Keywords: Adaptive estimation, Minimax procedures, Specification testing, Shrinkage, Robustness.

JEL classification codes: C13, C18.

*We thank Isaiah Andrews, Manuel Arellano, Dmitry Arkhangelsky, Stéphane Bonhomme, Tom Boot, Xu Cheng, Bryan Graham, Michal Kolesár, Roger Koenker, Lihua Lei, Jesse Shapiro, and Aleksey Tetenov for helpful discussions on this project. The paper also benefited from the comments of conference and seminar audiences. Timothy Armstrong gratefully acknowledges support from National Science Foundation Grant SES-2049765. Liyang Sun gratefully acknowledges support from the Institute of Education Sciences, U.S. Department of Education, through Grant R305D200010, and Ayudas Juan de la Cierva Formación. Code implementing the adaptive estimators proposed in this paper is available online at `https://github.com/lsun20/MissAdapt`.

†USC. Email: timothy.armstrong@usc.edu

‡UC Berkeley and NBER. Email: pkline@berkeley.edu

§UCL and CEMFI. Email: liyang.sun@ucl.ac.uk

1 Introduction

> Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful. – Box and Draper (1987)

Empirical research is typically characterized by a robustness-efficiency tradeoff. The researcher can either invoke strong assumptions to motivate an estimator that is precise, but sensitive to violations of model assumptions, or they can employ a less precise estimator that is robust to these violations. Familiar examples include the choice of whether to add a set of controls to a regression, whether to exploit over-identifying restrictions in estimation, and whether to allow for endogeneity or measurement error in an explanatory variable.

As the quote from Box and Draper illustrates, decisions of this nature are often approached with a degree of pragmatism: imposing a false restriction may be worthwhile if doing so yields improvements in precision that are not outweighed by corresponding increases in bias. While precision is readily assessed with asymptotic standard errors, the measurement of bias is less standardized. A popular informal approach is to conduct a series of "robustness exercises," whereby estimates from models that add or subtract assumptions from some baseline are reported and examined for differences. While robustness exercises of this nature can be informative, they can also be perplexing. How should the results of this exercise be used to refine the baseline estimate of the parameter of interest?

One answer, found often in econometrics textbooks, is to use a specification test to select a model. Doing so yields a *pre-test* estimator that equals the estimator of the restricted model when the specification test fails to reject, and is otherwise equal to the estimator of the unrestricted model. The pre-test estimator offers a form of asymptotic insurance against bias: as the degree of misspecification grows large relative to the noise in the data, the test rejects with near certainty. Yet when biases are modest, as one might expect of models that serve as useful approximations to the world, the cost of this insurance in terms of increased variance can be exceedingly high.

In this paper we explore an alternative to specification testing: *adapting* to misspecification.[1] Adaptive estimation provides a systematic approach to exploiting the assumptions of the restricted model as efficiently as possible while acknowledging the possibility that the

[1] An interactive Shiny application implementing our proposed estimator is available online at `https://lsun20.github.io/MissAdapt/`.

restriction in question is misspecified. Consider an oracle who knows a bound on the extent to which the restricted model is misspecified, allowing them to combine the estimates from the restricted and unrestricted models in a way that minimizes maximum risk. An adaptive estimator is one that comes as close as possible to achieving this oracle benchmark without using prior knowledge of the magnitude of misspecification.

We show that adaptive estimators can be computed by solving a weighted minimax problem. While the resulting *optimally adaptive* estimator does not have a closed form, an analytic soft-thresholding estimator can be tuned to yield comparable performance. This *adaptive soft-thresholding* estimator can be interpreted as a smoothed version of the pre-test estimator utilizing a critical value that depends on the correlation between the restricted and unrestricted estimators. The near-optimality of adaptive soft-thresholding contrasts with the performance of pre-test estimators, which perform poorly under moderate amounts of misspecification.

Both the optimally adaptive and adaptive soft-thresholding estimators are easily computed using information that is routinely reported in robustness checks. In the case where the restricted estimator is efficient under the restricted model, the estimators can be computed from published point estimates and standard errors alone. The adaptive soft-thresholding estimator can also be obtained via a particular sort of lasso regression (Tibshirani, 1996) that may be of independent interest in other low-dimensional settings.

To illustrate the advantages of adapting to—rather than testing for—misspecification, we revisit two empirical examples where questions of model specification arise. Our leading example, which we return to throughout the paper, is drawn from de Chaisemartin and D'Haultfœuille (2020b)'s reanalysis of Gentzkow et al. (2011), in which a two-way fixed effects estimator that exhibits negative weights in many periods is compared to a more variable convex weighted estimator. The optimally adaptive and adaptive soft-thresholding estimators are shown to place roughly equal weight on these two estimators. A second example, taken from Angrist and Krueger (1991), compares an ordinary least squares (OLS) estimate of the returns to schooling to an instrumental variables (IV) estimate. We argue that extra care is required in this example because the IV estimate is orders of magnitude less precise than OLS. Online Appendix E provides an additional example, drawn from LaLonde (1986), illustrating the problem of estimating the effects of job training using a mix of control groups whose credibility can be ranked ex-ante. In all of the above examples,

adapting between models is found to yield a more attractive balance between efficiency and robustness than selecting a single model via pre-testing, with the adaptive soft-thresholding estimator performing especially well.

Related literature. Our analysis builds on early contributions by Hodges and Lehmann (1952) and Bickel (1983, 1984) who consider families of robustness-efficiency tradeoffs defined over pairs of nested models. We extend this work by considering a continuum of models, indexed by different degrees of misspecification. Our framework also allows for more general parameter spaces indexed by a regularity parameter, however computational constraints limit us to low dimensional applications in practice.

A large statistics literature considers the problem of adaptation, defined as the search for an estimator that performs nearly as well as an oracle with additional knowledge of the data generating process. We focus on the case where proximity to oracle performance is measured in terms of the ratio of actual to oracle risk, which mirrors the definition used in Tsybakov (1998) and leads to simple risk guarantees and statements about relative efficiency. While the high dimensional statistics literature has mostly focused on asymptotic rates and constants, we study exact computation of quantities of interest in low dimensional settings.

A large literature considers Bayesian and empirical Bayesian schemes for either model selection or model averaging (Akaike, 1973; Mallows, 1973; Schwarz, 1978; Leamer, 1978; Hjort and Claeskens, 2003). The proposed adaptive estimator can be viewed as a Bayes estimator that utilizes a robust prior guaranteeing bounded influence of specification biases on risk. In contrast to empirical Bayesian proposals (e.g., Green and Strawderman, 1991; Hansen, 2007; Hansen and Racine, 2012; Cheng et al., 2019; Fessler and Kasy, 2019) our analysis considers a scalar estimand, which renders Stein style shrinkage arguments inapplicable.

de Chaisemartin and D'Haultfœuille (2020a) study empirical risk minimization in an analogous setting with a scalar parameter and misspecification. They derive a combination estimator that exhibits a maximum decrease in risk over the unrestricted estimator greater than its maximum increase in risk over the unrestricted estimator. Risk-limited variants of our adaptive estimators are shown to also satisfy this property.

2 An introductory example

In this section, we illustrate our proposal at a high level via an empirical example, postponing the details to later discussion. Gentzkow et al. (2011) studied the effects of newspapers on voter turnout in US presidential elections using a two-way fixed effects (TWFE) model estimated in first differences. de Chaisemartin and D'Haultfœuille (2020b) showed that in settings featuring staggered adoption, like the one studied by Gentzkow et al. (2011), TWFE estimators identify potentially non-convex combinations of average treatment effects over time and across adoption cohorts.

Convexity of the weights defining a causal estimand θ is generally agreed to be an important desideratum, guaranteeing that when treatment effects are of uniform sign, θ will also possess that sign. However, when treatment effect heterogeneity is mild, an estimator exhibiting asymptotic weights of mixed sign may yield negligible asymptotic bias and substantially lower asymptotic variance than a convex weighted alternative. Consequently, researchers choosing between traditional TWFE estimators and modern convex weighted alternatives often face a non-trivial robustness-efficiency tradeoff.

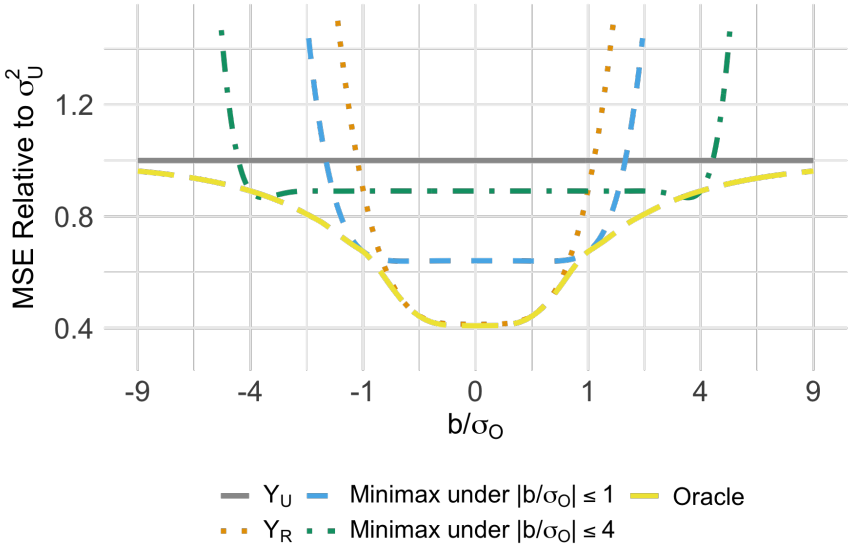
Let Y_R denote the first differenced TWFE estimator used by Gentzkow et al. (2011) and Y_U the convex weighted estimator proposed by de Chaisemartin and D'Haultfœuille (2020b). The restricted estimator Y_R evaluates to 0.26 – an additional newspaper raises voter turnout by 0.26 percentage points – with a standard error of $\sigma_R = 0.09$. The unrestricted estimator Y_U evaluates to 0.43 with a standard error of $\sigma_U = 0.14$. Suppose that Y_U and Y_R are normally distributed with standard deviations given by these standard errors, an approximation that can be formally justified using a local asymptotic misspecification framework.

Following de Chaisemartin and D'Haultfœuille (2020b), we assume that the target parameter θ is the average effect of an additional newspaper in counties gaining or losing a newspaper and that Y_U provides an unbiased estimator of this parameter. In contrast, the two-way fixed effects estimator will tend to identify a different parameter, yielding an unknown bias $b = E[Y_R] - \theta$. The difference $Y_O = Y_R - Y_U$ between the restricted and unrestricted estimators gives a noisy estimate of the bias b that forms the basis for standard "over-identification" tests of specification. To further simplify the example, suppose that $\text{cov}(Y_R, Y_O) = 0$. This condition, which seems to be very nearly satisfied in the data, implies that Y_R is efficient under the constraint $b = 0$. Consequently, the variance of Y_O is given by

$\sigma_O^2 = \sigma_U^2 - \sigma_R^2$ and the efficiency gain from imposing the restriction is given by σ_R^2/σ_U^2.

To compare these estimators, consider their mean squared error (MSE), which will be our preferred measure of risk. Since Y_U is unbiased, its MSE is equal to its variance $\sigma_U^2 = (0.14)^2$. In contrast, the MSE of the restricted estimator depends on its bias b: $E[(Y_R - \theta)^2] = b^2 + \sigma_R^2 = b^2 + (0.09)^2$. Figure 1 plots the MSE of the unrestricted and restricted estimators as functions of the unknown bias b. To ease visual interpretation both risk functions have been divided by $\text{var}(Y_U)$, which normalizes the risk of Y_U to 1.

Figure 1: Risk of unrestricted, restricted, B-minimax, and oracle estimators



Notes: Depiction assumes $\sigma_R^2/\sigma_U^2 = 0.41$. Horizontal axis is spaced quadratically.

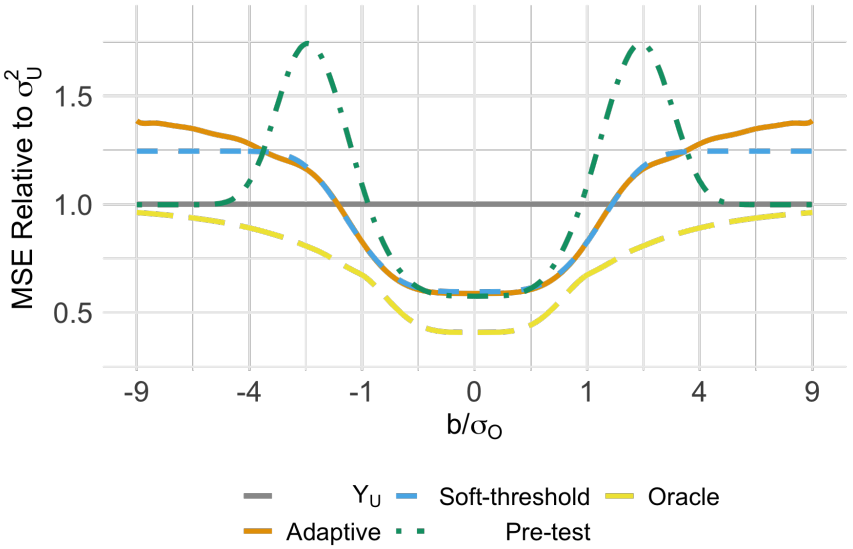
Reasonable people can disagree about which of these estimators is best. If $b = 0$, then Y_R gives a decrease in MSE from $(0.14)^2$ to $(0.09)^2$. The price paid for this improvement in MSE at $b = 0$ is that the MSE can be much larger than the MSE of the unrestricted estimator when $b \neq 0$. Of course, tradeoffs of this nature are unavoidable because Y_U is admissible: no other estimator has lower MSE for all b. The goal of adaptive estimation is to resolve this tradeoff by balancing efficiency when b is close to zero against robustness when b is large.

To find such an estimator, we introduce the benchmark of a hypothetical oracle that uses prior knowledge of a bound on the magnitude of the bias b to form an estimator. When $b = 0$, the oracle chooses Y_R, which is minimax when b is known to be 0. More generally, given a bound $B \geq 0$, one can compute the estimator that is minimax over the restricted parameter space $(\theta, b) \in \mathbb{R} \times [-B, B]$, a procedure we refer to as the *B-minimax estimator*. The oracle computes this estimator using prior knowledge of the best possible bound $B = |b|$, yielding

an *oracle estimator*. Though the oracle estimator lacks a closed form, it closely resembles a linear estimator $(1-w)Y_R + wY_U$ that uses prior knowledge of $|b|$ to choose an oracle weight $w = w_b^*$ that is decreasing in $|b|$. Figure 1 plots the risk function of the B-minimax estimator for $B \in \{1\sigma_O, 4\sigma_O\}$ along with the risk function of the oracle.

The oracle estimator cannot be computed without prior knowledge of the bias magnitude. One feasible stand-in for the oracle estimator is to posit a bound B on the bias and compute the B-minimax estimator. However, if this bound is set lower than the true bias magnitude $|b|$, we again expose ourselves to potentially very large MSE. An alternative to guessing a bound B is to use the data to infer a likely value of $|b|$. Then one can estimate θ optimally subject to the estimated bias magnitude. The pre-test estimator described in the introduction uses Y_U when $|Y_O| > 1.96\sigma_O$ and otherwise relies on Y_R. Pre-testing yields oracle-like behavior in some respects: when b is small, the restricted model tends to be selected, whereas, when b is large, the unrestricted model tends to be selected. Unfortunately, the risk function of the pre-test estimator, plotted in Figure 2, is quite large for moderate values of b, reflecting the cost of using the data "twice" in a non-smooth fashion.

Figure 2: Risk of optimally adaptive, soft thresholding, and pre test estimators



Notes: Depiction assumes $\sigma_R^2/\sigma_U^2 = 0.41$. Horizontal axis is spaced quadratically.

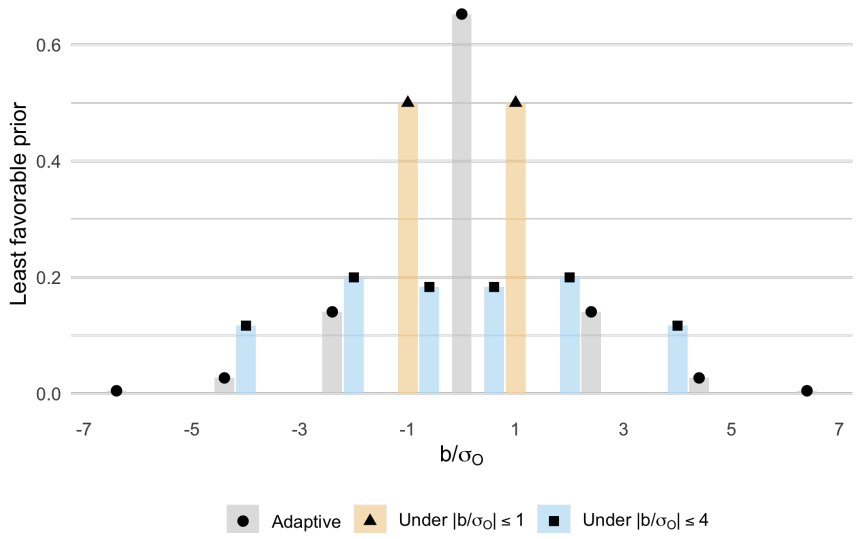
Adaptive estimators, by contrast, use the data to directly mimic the oracle's risk function. The *optimally adaptive estimator* is the estimator that comes closest to matching the oracle's risk function, where distance is measured in terms of the maximum ratio of actual to oracle risk across all bias levels, a metric that we term the *adaptation regret*.

While the optimally adaptive estimator lacks a closed form, a simple soft thresholding estimator can be tuned to come close to minimizing adaptation regret. Like the pre-test estimator, the *adaptive soft thresholding estimator* is equal to Y_R if $|Y_O/\sigma_O|$ is less than some threshold value λ. However, rather than switching discontinuously to Y_U when $|Y_O| > \lambda\sigma_O$, the soft thresholding estimator "shrinks" the unrestricted estimator towards the restricted estimator by λ standard errors of the bias estimate – i.e., the soft thresholding estimator equals $Y_U - \lambda\sigma_O$ when $Y_O < -\lambda\sigma_O$ and $Y_U + \lambda\sigma_O$ when $Y_O > \lambda\sigma_O$. The optimal threshold is a decreasing function of the ratio σ_R^2/σ_U^2, which captures the relative efficiency of Y_U to Y_R. In the present example, $\sigma_R^2/\sigma_U^2 = 0.41$, implying Y_U is only 41% as efficient as Y_R when $b = 0$. The optimal threshold in this case is $\lambda = 0.64$, far below the traditional 1.96 value used for pre-testing.

The risk function of the optimally adaptive estimator and its soft thresholding approximation are shown in Figure 2. The MSE of the optimally adaptive estimator is never more than 44% above the oracle MSE, which is the best that can be achieved. The adaptive soft thresholding estimator has an MSE that is never more than 46% above the oracle. When $b = 0$, these adaptive estimators achieve substantially (40% - 41%) lower MSE than Y_U. Conversely, when $|b|$ is large, they exhibit modestly (29% - 39%) higher MSE than Y_U. The pre-test estimator also achieves near oracle MSE levels when $b = 0$. However, when $|b| \approx 1.96\sigma_O$, its MSE is 118% percent above the oracle MSE and 75% above the MSE of Y_U. Evidently, the adaptive estimators yield both lower worst case departures from oracle risk and lower worst case MSE than pre-testing.

While the difficulty of eliciting prior beliefs about the bias b of the restricted estimator is one motivation for adaptation, both the adaptive estimator and its B-minimax counterparts can actually be thought of as Bayes estimators motivated by particular least favorable priors. Figure 3 depicts the least favorable priors utilized by the B-minimax estimator for two values of B along with the least favorable prior of the adaptive estimator. All three priors are discrete, symmetric about zero, and decreasing in $|b|$. Hence, all three estimators will tend to be more efficient than Y_U when the true bias magnitude $|b|$ is small. The adaptive prior has the important advantage over B-minimax priors of not requiring specification of the bound B. A second advantage of the adaptive prior is that it is *robust*: the risk of the optimally adaptive estimator remains bounded as $|b|$ grows large. In contrast, the risk of a B-minimax estimator grows rapidly and without limit once $|b|$ exceeds the posited bound B.

Figure 3: Least favorable priors when $\sigma_R^2/\sigma_U^2 = 0.41$



3 Setup

Consider a researcher who observes data or initial estimate Y taking values in a set $\mathcal{Y}$, following a distribution $P_{\theta,b}$ that depends on unknown parameters (θ, b). Let $E_{\theta,b}$ denote expectation under the distribution $P_{\theta,b}$. We will study possibly misspecified models in a normal or asymptotically normal setting. Results covering more general models are available in a prior version of this paper (Armstrong et al., 2023).

The random variable $Y = (Y_U, Y_R)$ consists of an unrestricted estimator Y_U of a scalar parameter $\theta \in \mathbb{R}$ and a restricted estimator Y_R that is predicated upon additional model assumptions. The additional restrictions required to motivate the restricted estimator make it less robust but potentially more efficient. To capture this tradeoff, we assume that Y_U is asymptotically unbiased for θ, while Y_R may exhibit a bias of b stemming from violation of the additional restrictions. We focus on the case where Y_R is a single scalar-valued estimate, but extensions to vector-valued b are provided in Appendix B.1.

It will often be convenient to work with the quantity $Y_O = Y_R - Y_U$, which gives an estimate of the bias b that features in conventional tests of over-identifying restrictions. We work with the large sample approximation

$$\begin{pmatrix} Y_U \\ Y_O \end{pmatrix} \sim N\left(\begin{pmatrix} \theta \\ b \end{pmatrix}, \Sigma\right), \quad \Sigma = \begin{pmatrix} \sigma_U^2 & \rho\sigma_U\sigma_O \\ \rho\sigma_U\sigma_O & \sigma_O^2 \end{pmatrix}. \tag{1}$$

The variance matrix Σ is treated as known. In practice, feasible versions of our procedures can be computed using a consistent estimate of the asymptotic variance matrix. The model (1) arises as from a local asymptotic framework where θ and b are scaled by the square root of the sample size and Y_U and Y_R are asymptotically normal. While we do not pursue formal statements translating our results about minimax and adaptive estimators to this framework, Armstrong and Kolesár (2021, Section 4.2 and Appendix C) obtain such results for efficiency bounds for CIs.

Under the restriction $b = 0$, the efficient estimator GMM estimator of θ is $Y_{R,GMM}$ and its variance is $\sigma^2_{R,GMM}$, where

$$Y_{R,GMM} := Y_U - (\rho\sigma_U/\sigma_O)Y_O, \quad \sigma^2_{R,GMM} := \text{var}(Y_{R,GMM}) = \sigma^2_U \cdot (1 - \rho^2). \tag{2}$$

In the case where $\rho\sigma_U\sigma_O = -\sigma^2_O$, the restricted estimator Y_R and the efficient GMM estimator $Y_{R,GMM}$ coincide because $\text{cov}(Y_R, Y_O) = 0$. One can compute σ^2_O in this case simply by subtracting the squared standard error of the restricted estimator from that of the unrestricted estimator (Hausman, 1978). The *relative efficiency* of Y_U to $Y_{R,GMM}$ is given by $\sigma^2_{R,GMM}/\sigma^2_U = 1 - \rho^2$.

3.1 B-minimax estimators

An estimator $\hat{\theta} : \mathcal{Y} \to \mathcal{A}$ maps the data Y to an action $a \in \mathcal{A}$. The loss of taking action a under parameters (θ, b) is given by the function $L(\theta, b, a)$. While it is possible to analyze many types of loss functions in our framework, we will focus on the familiar case of estimation of a scalar parameter $\theta \in \mathbb{R}$ with $\mathcal{A} = \mathbb{R}$ and squared error loss $L(\theta, b, \hat{\theta}) = (\hat{\theta} - \theta)^2$.

The risk of an estimator is given by the function

$$R(\theta, b, \hat{\theta}) = E_{\theta,b}L(\theta, b, \hat{\theta}(Y)) = \int L(\theta, b, \hat{\theta}(y))\, dP_{\theta,b}(y).$$

An estimator $\hat{\theta}$ is *minimax* over the set $\mathcal{C}$ for the parameter (θ, b) if it minimizes the maximum risk over $(\theta, b) \in \mathcal{C}$. We are interested in a setting where the researcher entertains multiple parameter spaces $\mathcal{C}_B$, indexed by $B \in \mathcal{B}$, which may restrict the parameters (θ, b) in different

ways. The maximum risk over the set $\mathcal{C}_B$ is

$$R_{\max}(B, \hat{\theta}) = \sup_{(\theta,b)\in\mathcal{C}_B} R(\theta, b, \hat{\theta}).$$

An estimator $\hat{\theta}$ is *minimax* over $\mathcal{C}_B$ if it minimizes $R(B, \hat{\theta})$. We denote such a "B-minimax" estimator by $\hat{\theta}^*_B$. The *minimax risk* for the parameter space $\mathcal{C}_B$ is the maximum risk of the minimax estimator:

$$R^*(B) = R_{\max}(B, \hat{\theta}^*_B) = \inf_{\hat{\theta}} R_{\max}(B, \hat{\theta}) = \inf_{\hat{\theta}} \sup_{(\theta,b)\in\mathcal{C}_B} R(\theta, b, \hat{\theta}).$$

We will refer to the quantity $R^*(B)$ as the *B-minimax risk*.

In our framework, the parameter space $\mathcal{C}_B$ is indexed by a scalar bound B on the magnitude of the bias of the restricted estimator:

$$\mathcal{C}_B = \{(\theta, b) : \theta \in \mathbb{R}, b \in [-B, B]\} = \mathbb{R} \times [-B, B].$$

Hence, the set $\mathcal{C}_\infty$ corresponds to the unrestricted parameter space, while $\mathcal{C}_0$ corresponds to the restricted parameter space. Consequently, the ∞-minimax estimator (the B-minimax estimator when $B = \infty$) is Y_U, while the 0-minimax estimator (the B-minimax estimator when B=0) is $Y_{R,GMM}$. In the special case where the restricted estimator is fully efficient, the 0-minimax estimator is additionally equal to the restricted estimator $Y_R = Y_U + Y_O$.

In some cases, researchers may have additional information about the problem at hand that motivates working with parameter spaces of a different nature. For instance, one might have ex-ante knowledge of the sign of the bias in Y_R. In such an example, Y_U would not be the ∞-minimax estimator under the relevant (sign restricted) $\mathcal{C}_B$. While the tools developed here to compute B-minimax and adaptive estimators are easily extended to other sorts of parameter spaces indexed by a scalar, we do not pursue such extensions in this paper.

3.2 Adaptation

B-minimax estimators provide a natural approach to incorporating prior restrictions into estimation. However, researchers are often unwilling to commit to a restricted parameter space $\mathcal{C}_B$, either because they lack appropriate prior information or because priors differ

among their scientific peers. In contrast to B-minimax estimators, adaptive estimators yield worst case risk near $R^*(B)$ for all B. That is, they yield uniformly "near-minimax" performance without commitment to a particular choice of B.

How close to B-minimax performance can one get without specifying B? Relative to an oracle that knows $|b| \leq B$ and is able to compute the B-minimax estimator $\hat{\theta}^*_B$, an estimator $\hat{\theta}$ formed without reference to a particular parameter space $\mathcal{C}_B$ yields a proportional increase in worst-case risk given by

$$A(B, \hat{\theta}) = \frac{R_{\max}(B, \hat{\theta})}{R^*(B)}.$$

We refer to $A(B, \hat{\theta})$ as the *adaptation regret* of the estimator $\hat{\theta}$ under the set $\mathcal{C}_B$. In our main results, risk corresponds to mean squared error. Hence, $(A(B, \hat{\theta}) - 1) \times 100$ gives the percentage increase in worst-case MSE over $\mathcal{C}_B$ faced by an estimator $\hat{\theta}$ relative to $\hat{\theta}^*_B$. Importantly, this regret notion is scale invariant: a change of the units in which MSE is measured (e.g., dollars squared versus squared cents) will not alter the percentage increase in risk over an oracle.

The adaptation regret may be as large as $A_{\max}(\mathcal{B}, \hat{\theta}) = \sup_{B \in \mathcal{B}} A(B, \hat{\theta})$, a quantity we term the *worst case adaptation regret*. The lowest possible value $A_{\max}(\mathcal{B}, \hat{\theta})$ can take is

$$A^*(\mathcal{B}) = \inf_{\hat{\theta}} \sup_{B \in \mathcal{B}} A(B, \hat{\theta}) = \inf_{\hat{\theta}} \sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \hat{\theta})}{R^*(B)}. \tag{3}$$

Following Tsybakov (1998), $A^*(\mathcal{B})$ gives the *loss of efficiency under adaptation*. An estimator $\hat{\theta}$ is *optimally adaptive* if $A_{\max}(\mathcal{B}, \hat{\theta}) = A^*(\mathcal{B})$. We use the notation $\hat{\theta}^*$ to denote such an estimator.

To measure the efficiency of an ad hoc estimator $\hat{\theta}$ relative to the optimally adaptive estimator, one can compute

$$\frac{A^*(\mathcal{B})}{A_{\max}(\mathcal{B}, \hat{\theta})} = \frac{\inf_{\tilde{\theta}} A_{\max}(\mathcal{B}, \tilde{\theta})}{A_{\max}(\mathcal{B}, \hat{\theta})}.$$

We refer to this quantity as the *adaptive efficiency* of the estimator $\hat{\theta}$. Note that $A(B, \hat{\theta})^{-1} = R^*(B)/R_{\max}(B, \hat{\theta})$ gives the relative efficiency of the estimator $\hat{\theta}$ to $\hat{\theta}^*_B$ under the parameter space $\mathcal{C}_B$. The optimally adaptive estimator $\hat{\theta}^*$ yields the best possible relative efficiency

that can be obtained simultaneously for all $B \in \mathcal{B}$. The loss of efficiency under adaptation gives the reciprocal of this best possible simultaneous relative efficiency.

We study the case where $\mathcal{C}_B = \mathbb{R} \times [-B, B]$ and take the set of values of B under consideration to be $\mathcal{B} = [0, \infty]$. Early work by Bickel (1984) considered adapting over the granular set $\mathcal{B}^{gran} = \{0, \infty\}$. Naturally, it is easier to adapt to the elements of the finite set $\mathcal{B}^{gran}$ than to the infinite set $\mathcal{B}$. Consequently, $A^*(\mathcal{B}^{gran}) \leq A^*(\mathcal{B})$. However, consideration of $\mathcal{B}^{gran}$ may leave efficiency gains on the table for $0 < b < \infty$ because $R^*(b) \leq R^*(\infty)$.

Bickel (1982) studied an asymptotic regime where $A(B, \hat{\theta}^*)$ tended to one, implying no asymptotic loss of efficiency under adaptation. By contrast, in the high-dimensional statistics literature, estimators typically exhibit non-negligible loss of efficiency under adaptation. For instance, the lasso achieves asymptotic MSE exceeding that of an oracle that knows the identity of the nonzero coefficients by a term that grows with the log of the number of regressors considered (Bühlmann and van de Geer, 2011, Ch. 6).

3.3 When is adaptation desirable?

The optimally adaptive estimator is designed for settings where researchers believe the bias in the restricted estimator is limited but nonetheless have difficulty committing to a particular bound $B < \infty$. Hence, like their minimax antecedents, adaptive estimators can be viewed as providing a convenient alternative to Bayesian estimation that avoids the requirement to fully specify a prior. Minimax decisionmaking has famously been axiomatized in terms of ambiguity aversion (Gilboa and Schmeidler, 1989; Schmeidler, 1989). Adaptation regret can likewise be interpreted as capturing the regret an ambiguity averse researcher feels over having exposed themselves to an unnecessarily high level of worst case risk.

A different sort of justification for minimax decisions—attributable to Savage (1954)—involves the potential of such decisions to foster consensus in settings where priors differ among members of a group. In Appendix A we develop a stylized extension of this argument that illustrates the ability of adaptive decisions to foster consensus among "committees" characterized by different sets of beliefs. In the model, each committee will agree to a B-minimax decision, choosing B to appease its most skeptical member. However, different committees prefer different values of $B \in \mathcal{B}$. When the loss of efficiency under adaptation $A^*(\mathcal{B})$ is not too large, the committees will agree to jointly follow the optimally adaptive decision because every committee can be compensated for the small increase in maximum

risk over their preferred B-minimax level.

Taking the committees to represent different camps of researchers, the model suggests adaptive estimation can help to forge consensus between researchers with varying beliefs about the suitability of different econometric models. In accord with the notion that the desirability of an optimally adaptive decision stems from its semblance to the relevant B-minimax decision, the prospects for achieving consensus decrease with the loss of efficiency under adaptation $A^*(\mathcal{B})$, which itself depends on the range of beliefs in $\mathcal{B}$.

When $A^*(\mathcal{B})$ is small, one can simply proceed with the optimally adaptive estimator and avoid arguments about whether the restricted model is appropriate. If it is large, then there are substantive tradeoffs in choosing B that cannot be avoided. Depending on the range of beliefs entertained by the scientific audience, adaptation may still be attractive when $A^*(\mathcal{B})$ is large. However, researchers may find it preferable to hedge against large biases in such settings by placing a constraint on worst-case risk, an extension we consider in Section 4.4.

4 Main results

This section derives the form of the optimally adaptive estimator in our setting. We begin by noting that the problem of computing adaptive estimators can be reduced to computing minimax estimators with a scaled loss function. We next use this insight along with invariance arguments to derive the form of the minimax and optimally adaptive estimators.

4.1 Adaptation as minimax with scaled loss

Plugging in the definition of $R_{\max}(B, \hat{\theta})$ along with $\mathcal{B} = [0, \infty]$ and $\mathcal{C}_B = \mathbb{R} \times [-B, B]$, the criterion that the optimally adaptive estimator $\hat{\theta}^*$ minimizes can be written

$$\sup_{B \in [0,\infty]} \frac{R_{\max}(B, \hat{\theta})}{R^*(B)} = \sup_{B \in [0,\infty]} \sup_{\theta \in \mathbb{R}, b \in [-B,B]} \frac{R(\theta, b, \hat{\theta})}{R^*(B)} = \sup_{(\theta,b) \in \mathbb{R}^2} \sup_{B \in [|b|,\infty]} \frac{R(\theta, b, \hat{\theta})}{R^*(B)}$$

where the last equality follows by noting that the double supremum on either side of this equality is over the same set of values of (B, θ, b). Since $R^*(B)$ is increasing in B, the inner supremum is taken at $B = |b|$, which gives the following lemma.

Lemma 4.1. *The loss of efficiency under adaptation (3) is given by*

$$\inf_{\hat{\theta}} \sup_{(\theta,b)\in\mathbb{R}^2} \omega(b)R(\theta,b,\hat{\theta}) \quad \textit{where} \quad \omega(b) = 1/R^*(|b|)$$

and an estimator $\hat{\theta}^*$ that achieves this infimum (if it exists) is optimally adaptive.

Lemma 4.1 shows that finding an optimally adaptive decision can be written as a minimax problem with a weighted version of the original loss function. In particular, $\hat{\theta}^*$ is found to minimize the maximum (over θ, b) of the objective $\omega(b)R(\theta,b,\hat{\theta}) = E_{\theta,b}\omega(b)L(\theta,b,\hat{\theta}(Y))$. Hence, the optimal adaptive estimator corresponds to a minimax estimator under the loss function $\omega(b)L(\theta,b,\hat{\theta}(Y))$.

4.2 Minimax and adaptive estimators

According to Lemma 4.1, computing adaptive estimators amounts to solving a weighted minimax problem. In our setting, we can further simplify this problem using invariance. We focus here on the case of squared error loss $L(\theta,b,\hat{\theta}) = (\theta-\hat{\theta})^2$. Appendix B.1 provides proofs of the results in this section and covers general loss functions for estimation of the form $L(\theta,b,\hat{\theta}) = \ell(\theta-\hat{\theta})$. It will be useful to transform the data to (Y_U, T_O), where $T_O = Y_O/\sigma_O$ is the t-statistic for a specification test of the null that $b = 0$. This representation is equivalent to our original setting because σ_O is known.

Applying invariance arguments and the Hunt-Stein theorem, it follows that the B-minimax estimator $\hat{\theta}^*_B(Y_U, T_O)$ takes the form

$$\hat{\theta}(Y_U, T_O) = \rho\sigma_U\delta\left(T_O\right) + Y_U - \rho\sigma_U T_O = \rho\sigma_U\delta\left(T_O\right) + Y_{R,GMM}, \tag{4}$$

where $Y_{R,GMM}$ is the efficient GMM estimator given in (2). To build some intuition for this estimator, note that if $b \neq 0$, then $Y_{R,GMM}$ will exhibit a bias of $-(\rho\sigma_U/\sigma_O)b$. The estimator in (4) subtracts from the GMM estimate a corresponding estimate $-\rho\sigma_U\delta\left(Y_O/\sigma_O\right)$ of this bias term.

The $\delta\left(T_O\right)$ employed by the B-minimax estimator can be shown to evaluate to the *bounded normal mean* estimator $\delta^{\mathrm{BNM}}\left(T_O; B/\sigma_O\right)$, where $\delta^{\mathrm{BNM}}(y;\tau)$ denotes the minimax estimator of $\vartheta \in \mathcal{C} = [-\tau,\tau]$ when $Y \sim N(\vartheta,1)$. The bounded normal mean problem has been studied extensively (as detailed in Lehmann and Casella, 1998, Section 9.7(i), p. 425) and

we detail its computation in Online Appendix D.3. For finite τ, the minimax estimator is the posterior mean against a least favorable prior. Figure 3 illustrates several such priors. When the interval is small, the least favorable prior concentrates at the two endpoints. For larger intervals, it concentrates at a finite number of points within $[-\tau, \tau]$ (Casella and Strawderman, 1981). For $\tau = \infty$, the minimax estimator is T_O.

The corresponding B-minimax risk is

$$R^*(B) = \rho^2\sigma_U^2 r^{\mathrm{BNM}}\left(B/\sigma_O\right) + \sigma_U^2 - \rho^2\sigma_U^2, \tag{5}$$

where $r^{\mathrm{BNM}}(\tau)$ denotes minimax risk in the bounded normal mean problem. This expression was used to construct the oracle risk curve displayed in Figures 1 and 2.

By Lemma 4.1, it suffices to compute the minimax estimator for θ under the scaled loss function $R^*(|b|)^{-1}(\theta - \hat{\theta})^2$ where $R^*(B)$ is given in (5). Invariance arguments can again be applied to show that the optimally adaptive estimator takes the same form as in (4), but with δ given by the estimator $\delta^*(t;\rho)$, which minimizes

$$\sup_{\tilde{b}\in\mathbb{R}} \frac{E_{T\sim N(\tilde{b},1)}(\delta(T) - \tilde{b})^2 + \rho^{-2} - 1}{r^{\mathrm{BNM}}(|\tilde{b}|) + \rho^{-2} - 1}. \tag{6}$$

The loss of efficiency under adaptation $A^*([0,\infty])$ is given by the minimized value of (6).

We summarize these results in the following theorem, which is proved in Appendix B.1.

Theorem 4.1. *Consider the model in (1) with parameter spaces $\mathcal{C}_B = \mathbb{R} \times [-B, B]$ for $B \in \mathcal{B} = [0,\infty]$ and squared error loss $L(\theta, b, d) = (d-\theta)^2$. The following results hold:*

(i) The B-minimax estimator takes the form in (4) with $\delta(\cdot)$ given by $\delta^{\mathrm{BNM}}(\cdot; B/\sigma_O)$ and the minimax risk $R^*(B)$ is given by (5).

(ii) An optimally adaptive estimator is given by (4) with $\delta(\cdot)$ given by a function $\delta^*(t;\rho)$ that minimizes (6).

(iii) The loss of efficiency under adaptation $A^*(\mathcal{B})$ in (3) is equal to

$$\inf_{\delta}\sup_{\tilde{b}\in\mathbb{R}} \frac{E_{T\sim N(\tilde{b},1)}(\delta(T) - \tilde{b})^2 + \rho^{-2} - 1}{r^{\mathrm{BNM}}(|\tilde{b}|) + \rho^{-2} - 1} = \sup_{\pi}\inf_{\delta} \int \frac{E_{T\sim N(\tilde{b},1)}(\delta(T) - \tilde{b})^2 + \rho^{-2} - 1}{r^{\mathrm{BNM}}(|\tilde{b}|) + \rho^{-2} - 1}\, d\pi(\tilde{b})$$

where the supremum is over all probability distributions π on $\mathbb{R}$.

Theorem 4.1(i) states that an oracle who knows the best bound $B = |b|$ will apply the estimator in (4) with δ given by the bounded normal mean estimator δ^{BNM}. One could alternatively consider an oracle with full knowledge of b that is required to entertain a restricted class of estimators. Magnus (2002) shows that such an oracle will choose a linear estimator $\delta(T_O) = w(b)T_O$ for some scalar weight $w(b)$ even if the class of estimators $\delta(T_O)$ can take the form $w(T_O, b)T_O$, where $w(T_O, b)$ is constrained to be symmetric in T_O, bounded between 0 and 1, nondecreasing in T_O on $[0, \infty)$, and to satisfy certain continuity conditions. Constraining the oracle to choose among linear rules in our setting would yield similar results, as minimax risk in the bounded normal means problem does not change much when attention is restricted to linear estimators (Donoho et al., 1990).

The class of estimators in (4) was also considered by Magnus and Durbin (1999) in the context of linear regression, albeit without the use of invariance arguments or a criterion such as minimax or adaptation regret. Our use of invariance to derive (4) mirrors the arguments of Bickel (1984) in the granular case where $\mathcal{B} = \{0, \infty\}$, although the characterization of the adaptive estimator in Bickel (1984) is different and uses the fact that adaptation is between two parameter spaces, rather than a continuum of parameter spaces.

4.2.1 Computation, least favorable prior and lookup table

According to Theorem 4.1, the optimally adaptive estimator $\delta^*(t; \rho)$ can be computed as the solution to a weighted minimax problem over the scaled bias $\tilde{b} = b/\sigma_O$. We use the characterization of minimax estimators as Bayes estimators under a least favorable prior. Following part (iii) of Theorem 4.1, the problem is solved numerically using a discrete approximation to the prior over $\tilde{b}$, similar to recent work in econometrics that has numerically computed solutions to minimax problems in other settings (e.g. Chamberlain, 2000; Elliott et al., 2015; Müller and Wang, 2019; Kline and Walters, 2021). The least favorable prior distributions reported in Figure 3 were computed using this approach. The invariance arguments used to derive (6) imply an independent flat prior for θ. See Online Appendix D for details.

To ease computation of the optimally adaptive estimator, we solved for the function $\delta^*(t; \rho)$ numerically at a grid of values of ρ. Tabulating these solutions yields a simple lookup table that allows rapid retrieval of (a spline interpolation of) the empirically relevant function $\delta^*(\cdot; \rho)$. We detail the construction of this lookup table in Online Appendix D.5. After evaluating this function at the realized T_O, the remaining computations take an analytic

closed form and can be evaluated nearly instantaneously.

4.2.2 Weighted average interpretation

One can write the estimator in (4) as a weighted average:

$$w(T_O) \cdot Y_U + (1 - w(T_O)) \cdot Y_{R,GMM}, \tag{7}$$

where $w(T_O) = \delta(T_O)/T_O$ is a data-dependent weight. The B-minimax estimator takes $\delta(\cdot)$ to be a minimax estimator that uses the constraint $|b| \leq B$ with known B, whereas the optimally adaptive estimator takes as $\delta(\cdot)$ an estimator engineered to adapt to different values of B in this constraint. We find numerically that the adaptive estimator "shrinks" T_O towards zero, leading the weight $\delta(T_O)/T_O$ to fall between zero and one for all values of ρ.

The data dependent nature of the weight $w(T_O)$ is clearly crucial for the robustness properties of the optimally adaptive estimator. As T_O grows large, less weight is placed on the optimal GMM estimator and more weight is placed on the unrestricted estimator Y_U. If one were to commit ex-ante to a fixed (i.e., non-stochastic) weight on Y_U below one, the worst-case risk of the procedure would become unbounded because the optimal GMM estimator can exhibit arbitrarily large bias. Consequently, worst case adaptation regret would also become unbounded.

4.2.3 Impossibility of consistently estimating the asymptotic distribution

The distribution of an estimator of the form (4) can be derived by noting that Y_{GMM} and T_O are independent, with $Y_{GMM} \sim N(\theta - b\rho\sigma_U/\sigma_O, \sigma_U^2(1-\rho^2))$ and $T_O \sim N(b/\sigma_O, 1)$. Let Z_1 and Z_2 denote independent $N(0,1)$ random variables. Substituting $T_O = Z_1 + b/\sigma_O$ and $Y_{GMM} = \sigma_U\sqrt{1-\rho^2}Z_2 + \theta - b\rho\sigma_U/\sigma_O$ into (4) and rearranging terms yields

$$\frac{\hat{\theta}(Y_U, T_O) - \theta}{\sigma_U} = \rho\left[\delta\left(Z_1 + \tilde{b}\right) - \tilde{b}\right] + \sqrt{1-\rho^2}Z_2, \quad \text{where} \quad \tilde{b} = b/\sigma_O. \tag{8}$$

This representation holds under the distribution for (Y_U, T_O) maintained in (1), which provides an asymptotic approximation under local misspecification. In this asymptotic regime, consistent estimators of ρ, σ_U and σ_O are available via the usual asymptotic variance formulas used in overidentification tests for GMM. In contrast, b gives the limit of the bias of

the restricted estimator divided by $\sqrt{n}$ and cannot be consistently estimated. Consequently, it is not possible to consistently estimate the asymptotic distribution of $\hat{\theta}(Y_U, T_O)$.

For example, the MSE of the estimator $\hat{\theta}(Y_U, T_O)$ is

$$\sigma_U^2 \left[\rho^2 r(b/\sigma_O; \delta(\cdot)) + 1 - \rho^2\right], \quad \text{where} \quad r(\tilde{b}; \delta(\cdot)) = E_{T\sim N(\tilde{b},1)}(\delta(T) - \tilde{b})^2.$$

Figures 1 and 2 of Section 2 plot this quantity as a function of $\tilde{b}$ with consistent estimates of ρ, σ_U and σ_O plugged in. However, $\tilde{b}$ itself cannot be consistently estimated. See Leeb and Pötscher (2005) for a discussion of these issues in the context of pre-test estimators.

4.2.4 Confidence Intervals

Using (8), one can obtain a $100 \cdot (1-\alpha)\%$ CI that is valid under the parameter space $\mathcal{C}_B = \mathbb{R} \times [-B, B]$ for (θ, b) by using a critical value $c_\alpha(\tilde{B}) = c_\alpha(\tilde{B}; \rho, \delta)$ solving

$$\inf \chi \quad \text{s.t.} \quad \sup_{\tilde{b}:|\tilde{b}|\leq \tilde{B}} P\left(\left|\rho\left[\delta\left(Z_1 + \tilde{b}\right) - \tilde{b}\right] + \sqrt{1-\rho^2} Z_2\right| > \chi\right) \leq \alpha. \tag{9}$$

This critical value can then be used to form the *fixed length confidence interval* (FLCI) $\left\{\hat{\theta}(Y_U, T_O) \pm \sigma_U c_\alpha(B/\sigma_O)\right\}$, which is centered at the estimator $\hat{\theta}(Y_U, T_O)$. To emphasize the dependence on the parameter space $\mathcal{C}_B$ under which coverage is guaranteed, we will refer to such intervals as B-FLCIs. For example, one can form the B-FLCI centered at the B-minimax estimator by using the critical value $c_\alpha(B/\sigma_U)$ for this estimator (i.e., for $\delta(\cdot) = \delta^{\text{BNM}}(\cdot; B/\sigma_U)$). Setting $B = \infty$, the ∞-FLCI centered at the ∞-minimax estimator is the usual CI centered at the unrestricted estimator: $\{Y_U \pm z_{1-\alpha/2}\sigma_U\}$. This CI turns out to be larger than the B-FLCI centered at the B-minimax estimator for finite B, reflecting its validity over the larger parameter space $b \in \mathbb{R}$.

Centering the B-FLCI at the B-minimax estimator requires specifying B. One would ideally like to automate the choice of B for CI construction using an *adaptive CI* that has length close to an infeasible $|b|$-FLCI, while maintaining coverage for all $b \in \mathbb{R}$. Unfortunately, it can be shown formally that adaptive CIs do not exist in our setting: any CI that is valid for all $b \in \mathbb{R}$ must have average length close to the length $2z_{1-\alpha/2}\sigma_U$ of the CI centered at Y_U, even if b happens to be close to zero (see Section 4 of Armstrong and Kolesár, 2021).

Despite these impossibility results, one can compute a B-FLCI centered at the adaptive

estimator by computing the critical value $c_\alpha(B/\sigma_O; \rho, \delta^*(\cdot;\rho))$ for the adaptive estimator. One approach is to report an adaptive estimator along with the critical values for a 0-FLCI and ∞-FLCI, thereby summarizing the range of critical values needed to guarantee coverage under different assumptions. When $|\rho|$ is large, the critical value for a 0-FLCI will be far below the usual 1.96 benchmark for a 95% test. Conversely, the corresponding critical value for a ∞-FLCI interval will be much larger than 1.96, reflecting the inherent tradeoffs involved in centering the CI around the adaptive estimator rather than the unbiased estimator. Cai and Low (2005) discuss analogous tradeoffs involving centering in the context of nonparametric estimation.

An alternate approach, which we explore in our main empirical example, is to construct a B-FLCI for some intermediate value of B and report both its worst and best case coverage. Researchers who are open to trading off some worst-case coverage for a shorter CI or enhanced best-case coverage might find an interval centered around an adaptive estimator, offering coverage (say) between 90% and 97%, more appealing than a longer interval centered around Y_U that consistently provides 95% coverage. This interval could also be preferable to a slightly shorter 90% CI centered around Y_U, as the additional 7 percentage points of potential coverage may be more valuable than a modest reduction in length.

4.3 Analytic adaptive estimators

While the optimally adaptive estimator is straightforward to compute via convex programming and is trivial to implement once the solution is tabulated, it lacks a simple closed form. To reduce the opacity of the procedure, one can replace the term $\delta(T_O)$ in (4) with an analytic approximation.

A natural choice of approximations for $\delta(T_O)$ is the class of *soft-thresholding* estimators, which are indexed by a threshold $\lambda \geq 0$ and given by

$$\delta_{S,\lambda}(T) = \max\{|T| - \lambda, 0\}\operatorname{sgn}(T) = \begin{cases} T - \lambda & \text{if } T > \lambda \\ T + \lambda & \text{if } T < -\lambda \\ 0 & \text{if } |T| \leq \lambda. \end{cases}$$

We also consider the class of *hard-thresholding* estimators, which are given by

$$\delta_{H,\lambda}(T) = T \cdot I(|T| \geq \lambda) = \begin{cases} T & \text{if } |T| > \lambda \\ 0 & \text{if } |T| \leq \lambda. \end{cases}$$

Note that hard-thresholding leads to a simple pre-test rule: use the unrestricted estimator if $|T_O| > \lambda$ (i.e. if we reject the null that $b = 0$ using critical value λ) and otherwise use the GMM estimator that is efficient under the restriction $b = 0$. The soft-thresholding estimator uses a similar idea, but avoids the discontinuity at $T_O = \lambda$.

As detailed in Appendix B.2, the soft-thresholding estimator is numerically equivalent to a generalized lasso estimator (Tibshirani, 1996) applied to a dataset comprised of the restricted and unrestricted estimates. The regressors are a constant and an indicator for the restricted estimate, the coefficient on which measures the bias b. The lasso penalty shrinks the bias estimate towards zero and depends only on the soft threshold λ. Hence, the adaptive soft threshold provides an optimal tuning of lasso for low-dimensional settings in which interest centers on a scalar parameter. This exact tuning contrasts with high-dimensional settings where existing tuning methods typically only offer rate results.

A third estimator, which we will call the empirical risk minimizer (ERM), takes the form $\delta_{ERM}(T_O) = \frac{T_O^2}{T_O^2+1} \cdot T_O$. The ERM estimator, which was proposed by de Chaisemartin and D'Haultfœuille (2020a), minimizes the estimated risk of the weighted average between Y_U and Y_{GMM}. The ERM can be generalized to a broader class of estimators

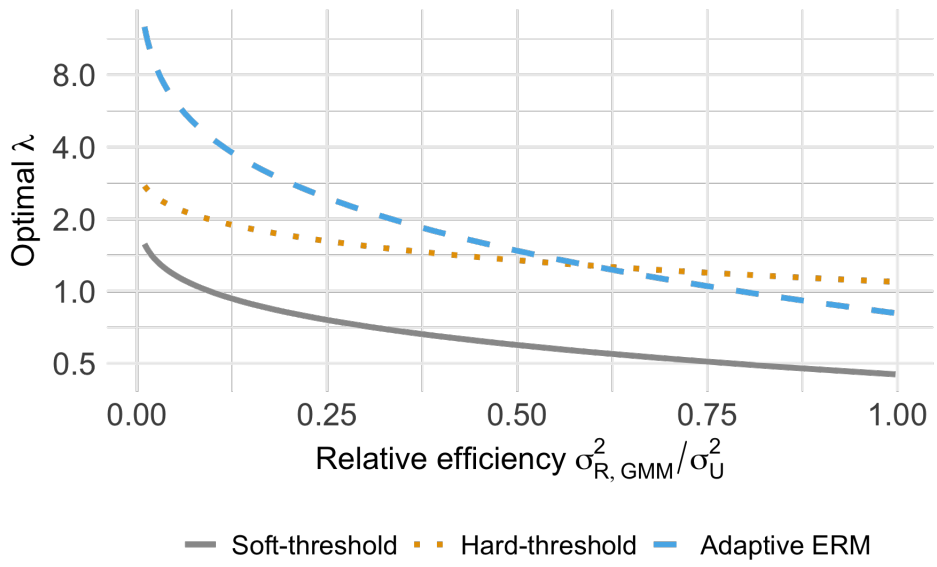
$$\delta_{ERM,\lambda}(T_O) = \frac{T_O^2}{T_O^2 + \lambda} \cdot T_O,$$

which was briefly considered in Magnus (2002, p. 230). We can optimize λ for the worst-case adaptation regret given a specific value of ρ^2, which yields the *adaptive ERM* estimator.

To compute the adaptive ERM estimator along with the hard and soft-thresholding estimators that are optimally adaptive in these classes of estimators, we minimize (6) numerically over λ. The minimax theorem does not apply to these restricted classes of estimators. Fortunately, the resulting two dimensional minimax problem in λ and $\tilde{b}$ is easily solved in practice as explained in Online Appendix D.6.

The optimized value of (6) then gives the worst-case adaptation regret of the adaptive

Figure 4: Thresholds minimizing the worst-case adaptation regret



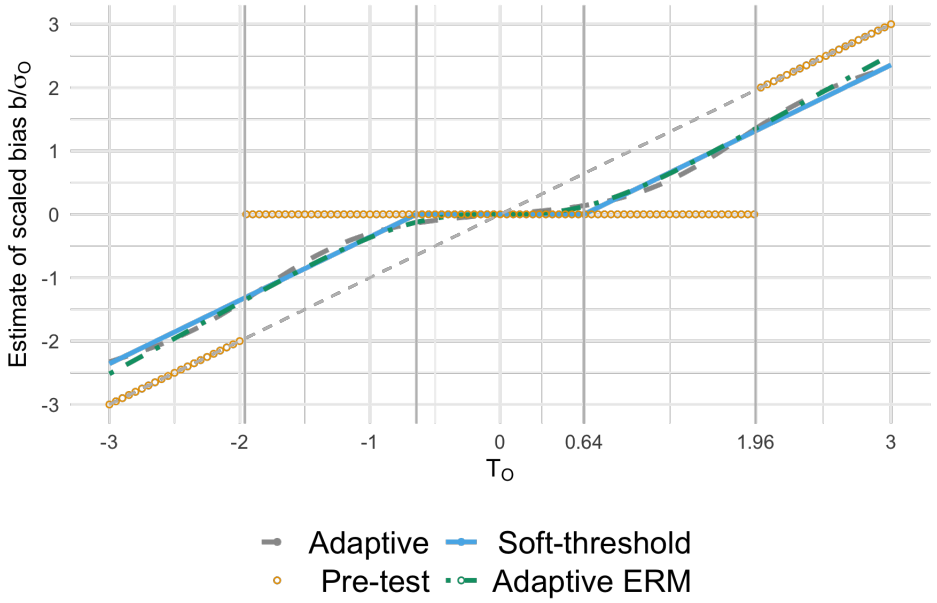
Notes: Vertical axis is spaced on a $\log_2$ scale.

ERM estimator or the adaptive soft or hard-thresholding estimator. We plot the respective optimal thresholds in Figure 4, which are only a function of the relative efficiency $\sigma^2_{R,GMM}/\sigma^2_U = 1-\rho^2$. We will be especially interested in the optimal soft threshold, which can be closely approximated using the formula $\lambda = 0.45 - 0.24 \cdot \ln(1-\rho^2)$ for $\rho^2 \in (0.002, 0.99)$.

Figure 5 plots the optimally adaptive and soft-thresholding estimators of the scaled bias as functions of T_O. To ease visual inspection of the differences between these estimators, they have been plotted over the restricted range [-3,3]. These functions depend on the data only through the estimated value of $1-\rho^2$, which takes the value 0.41 here, as in the two-way fixed effects example introduced in Section 2. The optimal soft-threshold λ yielding the lowest worst cast adaptation regret in this example is 0.64. The optimally adaptive, adaptive ERM, and soft-thresholding estimators continuously shrink small values of T_O towards zero. However, the soft-thresholding estimator sets all values of $|T_O|$ less than 0.64 to zero, while the optimally adaptive and adaptive ERM estimators avoid flat regions. In contrast to the continuous nature of these adaptive estimators, a conventional pre-test using $\lambda = 1.96$ exhibits large discontinuities at the hard threshold. The pre-test choice of $\lambda = 1.96$ differs from the value that minimizes worst-case adaptation regret, which in this example is 1.43.

Like the optimally adaptive estimator $\hat{\theta}^*$, the worst-case adaptation regret of the adap-

Figure 5: Estimators of scaled bias when $\sigma^2_{R,GMM}/\sigma^2_U = 0.41$



Notes: Solid vertical line at 0.64 depicts optimal soft-threshold. Solid line at 1.96 depicts conventional pre-test threshold.

tive soft and hard-thresholding estimators depends only on $1 - \rho^2$. We report comparisons between these estimators in our empirical applications in Section 5. As discussed in Online Appendix C.2, soft-thresholding yields nearly optimal performance for the adaptation problem relative to $\hat{\theta}^*$ in a wide range of settings. In contrast, hard-thresholding typically exhibits both substantially elevated worst case adaptation regret and worst case risk driven by the possibility that the scaled bias has magnitude near λ. The adaptive ERM estimator generally exhibits slightly higher worst case risk and adaptation regret than the soft-thresholding estimator but exhibits lower risk when the bias is very large.

Our finding that soft-thresholding is nearly optimal for adaptation mirrors the findings of Bickel (1984) for the case where the set $\mathcal{B}$ of bounds B on the bias consists of the two elements 0 and ∞. Magnus (2002, p. 231) reports that soft-thresholding (which he refers to as the Burr estimator) optimizes a related regret problem over a certain class of estimators indexed by two scalar parameters. While soft-thresholding is perhaps the simplest way of achieving near-optimal performance for adaptation, other generalizations of thresholding estimators (e.g., Johnstone, 2019, pp. 200-201) have been found to have similar risk properties to soft-thresholding, and may also perform well in our setting.

4.4 Constrained adaptation

If the loss of efficiency under adaptation $A^*(\mathcal{B})$ is large, both the optimally adaptive estimator and its soft-thresholding approximation will possess worst case risk far above the oracle minimax risk, which limits their practical appeal as devices for building consensus among researchers with different priors. As we show in Online Appendix C.3, $A^*(\mathcal{B})$ will tend to be large when $|\rho|$ is large, which corresponds to settings where Y_R is orders of magnitude more precise than Y_U. In such settings, substantial weight will be placed on the GMM estimator to guard against the immense adaptation regret that would emerge if $b = 0$, which exposes the researcher to severe biases if $|b|$ is large.

In such cases, it may be attractive to temper the degree of adaptation that takes place by restricting attention to estimators that exhibit worst case risk no greater than a constant $\bar{R}$. Online Appendix Section C.1 details how to compute such a constrained adaptive estimator. As noted by Bickel (1984) in his analysis of the granular case where $\mathcal{B} = \{0, \infty\}$, it is often possible to greatly improve the risk at $b = 0$ relative to the unbiased estimator Y_U in exchange for modest increases in risk when $b = \infty$. Similarly, we find that setting $\overline{R}$ to 50% above the risk of Y_U yields large efficiency improvements when b is small.

The constrained adaptive estimator bears some similarity to the ERM estimator. de Chaisemartin and D'Haultfœuille (2020a) prove that the maximal risk decrease of δ_{ERM} relative to the risk of the unbiased estimator is larger than the maximal risk increase of δ_{ERM} relative to the unbiased estimator. Through numerical calculations reported in a prior version of the paper (Armstrong et al., 2023), we find that this property holds for the constrained soft-thresholding version of our estimator so long as $\overline{R}$ is less than 70% above the risk of Y_U. Remarkably, the property holds even for unconstrained soft-thresholding ($\overline{R} = \infty$) so long as ρ^2 is less than 0.86.

5 Examples

We now consider two empirical examples where questions of specification arise and examine how adapting to misspecification compares to pre-testing and other strategies such as committing ex-ante to either the unrestricted or restricted estimator. Because the only inputs required to compute the adaptive estimator are the restricted and unrestricted point estimates along with their estimated covariance matrix, the burden on researchers of re-

porting adaptive estimates is very low. The analysis below draws on published tables of point estimates and standard errors whenever possible, using the replication data only to derive estimates of the covariance between the estimators. In both examples, we find that the restricted estimator is nearly efficient, implying the relevant covariances could have been inferred from published standard errors. A third example, provided in Appendix E, considers a multivariate adaptation problem with two restricted models and corresponding bias estimates.

5.1 Adapting to heterogeneous effects (Gentzkow et al., 2011)

Returning to the example introduced in Section 2, Gentzkow et al. (2011) study the effect of newspapers on voter turnout in US presidential elections between 1868 and 1928. They consider the following linear model relating the first-difference of the turnout rate to the first difference of the number of newspapers available in different counties:

$$\Delta y_{ct} = \beta \Delta n_{ct} + \Delta \gamma_{st} + \Delta \varepsilon_{ct},$$

where Δ is the first difference operator, y_{ct} is voter turnout per eligible voter in county c and year t, n_{ct} denotes the number of newspapers, and γ_{st} is a state by year fixed effect. The parameter β is meant to capture a causal effect of newspapers on voter turnout. In what follows, we take the OLS estimator of β as Y_R.

Studying this estimator in a heterogeneous treatment effects framework, de Chaisemartin and D'Haultfœuille (2020b) establish that Y_R yields a linear combination of average causal effects across different time periods and different counties, estimating that 46% of the relevant combination weights are negative. To guard against the potential biases stemming from reliance on negative weights, they propose a convex weighted estimator of average treatment effects featuring weights that are treatment shares. We take this convex weighted estimator as Y_U, implying our estimand of interest θ is the average effect of a change in the number of newspapers on turnout in county-years where the number of newspapers changed. When treatment effects are constant, the two-way fixed effects estimator is also consistent for this parameter.

Estimates Table 1 reports the realizations of (Y_U, Y_R) and their standard errors, which exactly replicate those given in Table 3 of de Chaisemartin and D'Haultfœuille (2020b) after

dividing by 100. The estimated variance of Y_O is closely approximated by the difference in squared standard errors between Y_U and Y_R, suggesting Y_R is nearly efficient. Hence, the downstream GMM, adaptive, and soft-thresholding estimators could have been accurately approximated using only the published point estimates and standard errors. Standard errors are not reported for the soft-thresholding, adaptive, or pre-test estimators because the variability of these procedures depends on the unknown bias level b.

Table 1: Estimates of the effect of an additional newspaper on turnout.

	Y_U	Y_R	$Y_{R,GMM}$	Pre-test	Opt. Adapt	Soft-thresh	Hard-thresh	ERM	Adapt ERM
Estimate	0.43	0.26	0.24	0.24	0.36	0.36	0.43	0.38	0.36
Std Error	(0.14)	(0.09)	(0.09)						
Max Risk	0%	∞	∞	87%	39%	25%	39%	15%	25%
Max Regret	145%	∞	∞	134%	44%	46%	82%	68%	50%
Threshold				1.96		0.64	1.43	1	1.73

Notes: Bootstrap standard errors in parentheses computed using the same 100 bootstrap samples utilized by de Chaisemartin and D'Haultfœuille (2020b). The over-identification test statistic is $T_O = -1.75$. "Pre-test" selects between Y_U and GMM based on $|T_O| \gtrless 1.96\sigma_O$. The relative efficiency of Y_U to $Y_{R,GMM}$ is $1 - \rho^2 = 0.41$. "Max Risk" gives the percentage increase in worst case risk over Y_U: $(\sup_B R_{\max}(B, \hat{\theta})/\sigma_U^2 - 1) \times 100$. "Max Regret" refers to the worst case adaptation regret in percentage terms $(A_{\max}(\mathcal{B}, \hat{\theta}) - 1) \times 100$.

Though the realized value of Y_U is nearly twice as large as that of Y_R, the two estimators are not statistically distinguishable from one another at the 5% level. Hence, a conventional pre-test suggests ignoring the perils of negative weights and confining attention to Y_R on account of its substantially increased precision. The worst case MSE of the pre-test estimator is 75% higher than the MSE σ_U^2 of Y_U, reflecting the hump shaped risk profiles depicted in Figure 2. Pre-testing also yields sizable worst-case adaptation regret reflecting the possibility that the test selects the inefficient Y_U when $b = 0$. Like Y_R, $Y_{R,GMM}$ exhibits a standard error roughly 35% below that of Y_U. Consequently, relying solely on the convex-weighted (but highly inefficient) estimator Y_U exposes the researcher to a large worst-case adaptation regret of 145%.

In contrast to the pre-test estimator, both the optimally adaptive estimator and its soft-thresholding approximation place substantial weight $w(T_O)$ on the convex estimator, yielding estimates roughly 60% of the way towards Y_U from $Y_{R,GMM}$. This phenomenon owes to the fact that with $T_O = -1.7$ both estimators detect the presence of a non-trivial amount of bias in Y_R. We can easily compute the soft-thresholding bias estimate from the figures reported

in the table as $(-1.7 + .64) \times 0.1 \approx -.11$, suggesting that Y_R exhibits a bias of roughly 40%. Balancing this bias against the estimator's increased precision leads the soft-thresholding estimator to essentially split the difference between the convex and non-convex weighted estimators.

By construction, the adaptive estimator exhibits lower worst case adaptation regret than the soft-thresholding estimator but the differences are quantitatively trivial. However, the soft-thresholding estimator exhibits meaningfully lower worst case risk than the adaptive estimator. Though the two estimators happen to yield identical estimates ex-post in this example, the ex-ante risk properties of the adaptive soft-thresholding estimator arguably commend it over the optimally adaptive estimator.

The ERM estimator of de Chaisemartin and D'Haultfœuille (2020a) yields lower worst case risk than soft-thresholding but substantially larger adaptation regret. Optimizing the ERM threshold to minimize adaptation regret yields worst case risk equivalent to the soft-thresholding estimator but higher adaptation regret. Evidently, soft-thresholding offers the most attractive tradeoff between worst case risk and adaptation regret of the estimators considered.

Confidence Intervals Table 2 reports the best case and worst case coverage of a series of confidence intervals. The first two columns of Panel A show that the usual 95% confidence interval centered around the unbiased estimator has proper size, while a naive CI centered around the restricted estimator has best case coverage of 95% and worst case coverage of 0% attributable to the potentially unlimited bias of the restricted estimator. Relying on a pre-test to select one of these two confidence intervals yields a minimum coverage level of 67%. By contrast, centering a CI around the optimally adaptive estimator using the standard error of the unbiased estimator yields best case coverage of 98% and worst case coverage of 90%. Centering around the soft thresholding estimator yields even more favorable results, raising the worst case coverage to 93%.

Panel B of Table 2 considers B-FLCIs centered around the adaptive estimators. A 0-FLCI centered around the optimally adaptive estimator has a half length of only about $1.54\sigma_U$ (as opposed to the traditional $1.96\sigma_U$ utilized in Panel A) but exhibits worst case coverage of 80%. Centering around the soft thresholding estimator yields a slightly longer interval, which improves minimum coverage to 87%. The third row of Panel B shows the coverage of a σ_O-FLCI centered around the optimally adaptive estimator, which yields modestly longer CI

Table 2: Coverage and length of confidence intervals

Panel A: Simple CIs

	Y_U $\pm 1.96\sigma_U$	Y_R $\pm 1.96\sigma_R$	Pre-test	Opt. Adapt $\pm 1.96\sigma_U$	Soft-Thresh $\pm 1.96\sigma_U$
Max Coverage	95%	95%	95%	98%	98%
Min Coverage	95%	0%	67%	90%	93%

Panel B: B-FLCIs

	Opt. Adapt $\pm c_{.05}(0)\sigma_U$	Soft-Thresh $\pm c_{.05}(0)\sigma_U$	Opt. Adapt $\pm c_{.05}(1)\sigma_U$	Soft-Thresh $\pm c_{.05}(1)\sigma_U$	Opt. Adapt $\pm c_{.05}(9)\sigma_U$	Soft-Thresh $\pm c_{.05}(9)\sigma_U$
Max Coverage	95%	95%	97%	97%	99%	99%
Min Coverage	80%	87%	86%	90%	95%	95%
Critical Val	1.54	1.62	1.74	1.77	2.32	2.11

Notes: "Max coverage" refers to the maximal coverage probability for the given confidence interval. "Min Coverage" refers to the min coverage probability. "Adaptive" refers to the optimally adaptive estimator and "Soft-Thresh" refers to soft thresholding. "Pre-test" switches between $Y_U \pm 1.96\sigma_U$ and $Y_R \pm 1.96\sigma_R$ based on whether $|T_O| \gtrless 1.96\sigma_O$. Critical values for B-FLCIs found by solving (9). Min/max coverage evaluated using the expression for the constraint in (9).

but lowers worst case coverage to 86%. Again, centering at the soft thresholding estimator raises worst case coverage slightly, in this case to 90%. Finally, we approximate an ∞-FLCI by setting $B = 9\sigma_O$, which yields very conservative intervals with half-lengths exceeding $2.1\sigma_U$.

The simplicity and robustness of intervals based upon an adaptive estimator $\pm 1.96\sigma_U$ make them an attractive option. For researchers who seek shorter intervals, the σ_O-FLCI centered around the soft thresholding estimator seems to offer a reasonable mix of worst and best case coverage. Notably, all of these options offer substantially higher worst case coverage than pre-testing, which remains widespread in applied research.

5.2 Adapting to endogeneity (Angrist and Krueger, 1991)

Our second example, which is meant to highlight the limits of optimal adaptation, comes from Angrist and Krueger (1991)'s classic analysis of the returns to schooling using quarter of birth as an instrument for schooling attainment. Documenting that individuals born in the first quarter of the year acquire fewer years of schooling than those born later in the year, they demonstrate that those born in the first quarter of the year also earn less than

those born later in the year.

Table 3 replicates exactly the estimates reported in Angrist and Krueger (1991, Panel B, Table III) for men born 1930-39. Y_U gives the Wald-IV estimate of the returns to schooling using an indicator for being born in the first quarter of the year as an instrument for years of schooling completed, while Y_R gives the corresponding OLS estimate. Neither estimator controls for additional covariates. When viewed through the lens of the linear constant coefficient models that dominated labor economics research at the time, the IV estimator identifies the same parameter as OLS under strictly weaker exogeneity requirements. In particular, IV guards against "ability bias," which plagues OLS in such models (Griliches and Mason, 1972; Ashenfelter and Krueger, 1994).

The first stage relationship between quarter of birth and years of schooling exhibits a z-score of 8.24, suggesting an asymptotic normal approximation to Y_U is likely to be highly accurate. We follow the original study in assuming homoscedasticity, in which case OLS (Y_R) is known to be the asymptotically efficient GMM estimator under exogeneity.

Table 3: Estimates of the return to an additional year of schooling.

				Unconstrained			Constrained	
	Y_U	Y_R	Pre-test	Opt. Adapt	Soft-thresh	Hard-thresh	Opt. Adapt	Soft-thresh
Estimate	0.102	0.071	0.071	0.071	0.071	0.071	0.080	0.085
Std Error	(0.0239)	(0.0003)						
Max Risk	0%	∞	147%	455%	427%	608%	50%	50%
Max Regret	634,577%	∞	17,882%	493%	537%	709%	15,134%	17,926%
Threshold			1.96		2.07	3.30		0.71

Notes: Standard errors in parentheses computed under homoscedasticity as in original study. Under homoscedasticity, Y_R coincides with GMM. The over-identification test statistic is $T_O = -1.3$. "Max Risk" gives the percentage increase in worst case risk over Y_U: $(\sup_B R_{\max}(B, \hat{\theta})/\sigma_U^2 - 1) \times 100$. "Max regret" refers to the worst case adaptation regret in percentage terms $(A_{\max}(\mathcal{B}, \hat{\theta}) - 1) \times 100$. The relative efficiency of Y_U to $Y_R = Y_{R,GMM}$ is $1 - \rho^2 = 0.0004$.

While the IV estimator accounts for endogeneity, it is highly imprecise, with a standard error two orders of magnitude greater than OLS. Consequently, the maximal regret associated with using IV instead of OLS is extremely large, as Y_U is only 0.04% as efficient as Y_R when exogeneity holds. IV and OLS cannot be statistically distinguished at conventional significance levels, with $T_O \approx -1.3$. The inability to distinguish IV from OLS estimates of the returns to schooling is characteristic not only of the specifications reported in Angrist and Krueger (1991) but of the broader quasi-experimental literature spawned by their landmark

study (Card, 1999).

The confluence of extremely large maximal regret for Y_U with a statistically insignificant difference Y_O, leads the adaptive estimator, the soft-thresholding estimator, and the pre-test estimator to all coincide with Y_R. The motives for this coincidence are of course quite different. The adaptive and soft-thresholding estimators seek to avoid the regret associated with missing out on the enormous efficiency gains of OLS if it is unconfounded. By contrast, the pre-test estimator simply fails to reject the null hypothesis that years of schooling is exogenous at the proper significance level.

Despite the agreement of the three approaches, the extremely large adaptation regret exhibited by the optimally adaptive estimator suggests it is unlikely to garner consensus in this setting. Committing to Y_R exposes the researcher to potentially unlimited risk. The adaptive and soft-thresholding estimators avoid committing to either Y_U or Y_R before observing the data but still expose the researcher to more than a 400% increase in maximal risk over Y_U. A skeptic concerned with the potential biases in OLS is therefore unlikely to be willing to rely on such an estimator.

If we instead limit ourselves to a 50% increase in maximal risk, the adaptive and soft-threshold estimators yield returns to schooling estimates of 0.080 and 0.085 respectively. While the former estimate is a bit closer to OLS than IV, the latter is approximately halfway between the two. The maximal regret of both these estimators is extremely high, reflecting the potential efficiency costs of weighting Y_U so heavily. These efficiency concerns are likely outweighed in this case by the potential for extremely large biases. Though these estimates are unlikely to garner consensus across camps of researchers with widely different beliefs, the risk-limited adaptive estimators should yield wider consensus than proposals to discard the OLS estimates and rely on IV alone.

6 Conclusion

Empiricists routinely encounter robustness-efficiency tradeoffs. The reporting of estimates from different models has emerged as a best practice at leading journals. The methods introduced here provide a scientific means of summarizing what has been learned from such exercises and arriving at a preferred estimate that trades off considerations of bias against variance.

Computing the adaptive estimators proposed in this paper requires only point estimates, standard errors, and the covariance between estimators, objects that are easily produced by standard statistical packages. As our examples revealed, in many cases the restricted estimator is nearly efficient, implying the relevant covariance can be deduced from the standard errors of the restricted and unrestricted estimators.

In line with earlier results from Bickel (1984), we found that soft-thresholding estimators closely approximate the optimally adaptive estimator in the scalar case, while requiring less effort to compute. An interesting topic for future research is whether similar approximations can be developed for higher dimensional settings where the curse of dimensionality renders direct computation of optimally adaptive estimators infeasible.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd International Symposium on Information Theory, 1973*, pp. 267–281. Akademiai Kiado.

Angrist, J. D. and A. B. Krueger (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics 106*(4), 979–1014.

Armstrong, T. B., P. Kline, and L. Sun (2023). Adapting to misspecification. *arXiv preprint arXiv:2305.14265v3*.

Armstrong, T. B. and M. Kolesár (2021). Sensitivity analysis using approximate moment condition models. *Quantitative Economics 12*(1), 77–108.

Ashenfelter, O. and A. Krueger (1994). Estimates of the economic return to schooling from a new sample of twins. *The American economic review*, 1157–1173.

Bickel, P. J. (1982, September). On Adaptive Estimation. *The Annals of Statistics 10*(3), 647–671.

Bickel, P. J. (1983). Minimax estimation of the mean of a normal distribution subject to doing well at a point. In M. H. Rizvi, J. S. Rustagi, and D. Siegmund (Eds.), *Recent Advances in Statistics*, pp. 511–528. Academic Press.

Bickel, P. J. (1984, September). Parametric Robustness: Small Biases can be Worthwhile. *The Annals of Statistics 12*(3), 864–879. Publisher: Institute of Mathematical Statistics.

Box, G. E. and N. R. Draper (1987). *Empirical model-building and response surfaces.* John Wiley & Sons.

Bühlmann, P. and S. van de Geer (2011, June). *Statistics for High-Dimensional Data: Methods, Theory and Applications* (2011 edition ed.). Heidelberg ; New York: Springer.

Cai, T. T. and M. G. Low (2005, April). Adaptive estimation of linear functionals under different performance measures. *Bernoulli 11*(2), 341–358.

Card, D. (1999). The causal effect of education on earnings. *Handbook of labor economics 3*, 1801–1863.

Casella, G. and W. E. Strawderman (1981). Estimating a Bounded Normal Mean. *The Annals of Statistics 9*(4), 870 – 878.

Chamberlain, G. (2000, November). Econometric applications of maxmin expected utility. *Journal of Applied Econometrics 15*(6), 625–644.

Cheng, X., Z. Liao, and R. Shi (2019). On uniform asymptotic risk of averaging GMM estimators. *Quantitative Economics 10*(3), 931–979.

de Chaisemartin, C. and X. D'Haultfœuille (2020a, June). Empirical MSE Minimization to Estimate a Scalar Parameter.

de Chaisemartin, C. and X. D'Haultfœuille (2020b, September). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review 110*(9), 2964–2996.

Donoho, D. L. (1994, March). Statistical Estimation and Optimal Recovery. *The Annals of Statistics 22*(1), 238–270.

Donoho, D. L., R. C. Liu, and B. MacGibbon (1990, September). Minimax Risk Over Hyperrectangles, and Implications. *The Annals of Statistics 18*(3), 1416–1437.

Elliott, G., U. K. Müller, and M. W. Watson (2015, March). Nearly Optimal Tests When a Nuisance Parameter Is Present Under the Null Hypothesis. *Econometrica 83*(2), 771–811.

Fessler, P. and M. Kasy (2019). How to use economic theory to improve estimators: Shrinking toward theoretical restrictions. *Review of Economics and Statistics 101*(4), 681–698.

Gentzkow, M., J. M. Shapiro, and M. Sinkinson (2011, December). The Effect of Newspaper Entry and Exit on Electoral Politics. *American Economic Review 101*(7), 2980–3018.

Gilboa, I. and D. Schmeidler (1989). Maxmin expected utility with non-unique prior. *Journal of mathematical economics 18*(2), 141–153.

Green, E. J. and W. E. Strawderman (1991). A james-stein type estimator for combining unbiased and possibly biased estimators. *Journal of the American Statistical Association 86*(416), 1001–1006.

Griliches, Z. and W. M. Mason (1972). Education, income, and ability. *Journal of political Economy 80*(3, Part 2), S74–S103.

Hansen, B. E. (2007). Least Squares Model Averaging. *Econometrica 75*(4), 1175–1189.

Hansen, B. E. and J. S. Racine (2012). Jackknife model averaging. *Journal of Econometrics 167*(1), 38–46.

Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica 46*(6), 1251–1271.

Hjort, N. L. and G. Claeskens (2003). Frequentist Model Average Estimators. *Journal of the American Statistical Association 98*(464), 879–899.

Hodges, J. L. and E. L. Lehmann (1952). The use of Previous Experience in Reaching Statistical Decisions. *The Annals of Mathematical Statistics 23*(3), 396–407.

Johnstone, I. M. (2019). *Gaussian estimation: Sequence and wavelet models.* Online manuscript available at https://imjohnstone.su.domains/.

Kline, P. and C. Walters (2021). Reasonable doubt: Experimental detection of job-level employment discrimination. *Econometrica 89*(2), 765–792.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, 604–620.

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*, Volume 53. John Wiley & Sons Incorporated.

Leeb, H. and B. M. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*(01), 21–59.

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation* (2nd edition ed.). New York: Springer.

Magnus, J. R. (2002). Estimation of the mean of a univariate normal distribution with known variance. *The Econometrics Journal 5*(1), 225–236.

Magnus, J. R. and J. Durbin (1999). Estimation of Regression Coefficients of Interest when Other Regression Coefficients are of no Interest. *Econometrica 67*(3), 639–643.

Mallows, C. L. (1973). Some Comments on CP. *Technometrics 15*(4), 661–675.

Müller, U. K. and Y. Wang (2019, March). Nearly weighted risk minimal unbiased estimation. *Journal of Econometrics 209*(1), 18–34.

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley & Sons.

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica: Journal of the Econometric Society*, 571–587.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics 6*(2), 461–464.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology 58*(1), 267–288.

Tsybakov, A. B. (1998, December). Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *The Annals of Statistics 26*(6), 2420–2469.

Appendix A Group decision making interpretation

This appendix develops a stylized model of group decision making inspired by Savage (1954)'s arguments regarding the ability of minimax decisions to foster consensus among individuals

with heterogeneous beliefs. Extending these arguments, we illustrate how adaptive decisions can serve to foster consensus across groups of individuals with different sets of beliefs.

A.1 Consensus in a single committee

Suppose there is a committee charged with deciding on the value of a parameter θ (e.g., the social cost of carbon) based on the evidence (Y_U, Y_R). The committee is comprised of members with heterogeneous beliefs over (θ, b) that include all priors supported on the set $\mathcal{C}_B$. The committee chair, who we will call the *B-chair*, offers a take it or leave it proposal that her committee agree on the estimator $\hat{\theta}$ in exchange for the provision of a public good providing payoff G to each member of the committee. This public good might entail, for example, a reduction in committee work or an offer to end the meeting early.

If the committee agrees to the proposal, the B-chair earns a payoff $K - C(G)$, where K is the value of consensus and $C(\cdot)$ is an increasing cost function. If some member of the committee does not agree to the proposal, the chair and all committee members receive payoff zero. The B-chair therefore seeks an estimator $\hat{\theta}$ allowing payment of the smallest G that ensures consensus.

A committee member who is certain of the parameters (θ, b) will accept the chair's offer if and only if $R\left(\theta, b, \hat{\theta}\right) \leq G$. However, the committee member with the most pessimistic beliefs regarding these parameters will require a public goods provision level of at least $R_{\max}\left(B, \hat{\theta}\right)$ to agree to the offer. To achieve consensus at minimal cost, the B-chair can propose the B-minimax estimator $\hat{\theta}^*_B$, which requires public goods provision level $R^*(B)$ to achieve consensus.

The B-chair will be willing to provide this level of public goods if and only if $K \geq C(R^*(B))$, in which case consensus ensues. If this condition does not hold, the chair deems the B-minimax estimator too costly to implement and consensus is not achieved. Hence, when no individual holds beliefs that are too extreme, the minimax estimator fosters consensus.

A.2 Consensus among committees

Now suppose there is a collection $\mathcal{B}$ of committees (e.g., the Intergovernmental Panel on Climate Change), each of which must decide on the parameter θ using (Y_U, Y_R). This collection

is led by a *chair of chairs* (CoC) who would like for the B-chairs to agree on a common estimator $\hat{\theta}$. Suppose also that $K > \sup_{B \in \mathcal{B}} C(R^*(B))$, so that each B-chair would privately prefer the B-minimax estimator $\hat{\theta}^*_B$. The CoC has a fixed budget $F > 0$ that can be used to provide a public good $\tilde{G}$ enjoyed by all chairs. The CoC makes provision of $\tilde{G}$ contingent on the agreement of all B-chairs to use $\hat{\theta}$: if they fail to reach consensus, the public good is not provided. The cost to the CoC of providing public goods level $\tilde{G}$ is $\tilde{C}(\tilde{G})$, where $\tilde{C}(\cdot)$ is monotone increasing.

By the arguments above, each B-chair must pay a cost $C(R_{\max}\left(B, \hat{\theta}\right))$ to secure consensus regarding the CoC's proposed $\hat{\theta}$, leaving her with payoff $K - C(R_{\max}\left(B, \hat{\theta}\right))$. However, each chair can also defy the CoC and propose $\hat{\theta}^*_B$ to her committee, yielding payoff $K - C(R^*(B))$. Hence, to compel a B-chair to use $\hat{\theta}$, the CoC must offer a public good providing utility of at least $\Delta_B(\hat{\theta}) = C(R_{\max}\left(B, \hat{\theta}\right)) - C(R^*(B))$. To minimize costs, the CoC sets $\tilde{G} = \sup_{B \in \mathcal{B}} \Delta_B(\hat{\theta})$, which is the smallest level capable of appeasing the most reticent B-chair.

Different functional forms for the cost function C yield different notions of adaptation. To motivate the formulation in (3), we assume $C(G) \propto \ln G$, which implies chairs produce the public good according to an increasing returns to scale technology that is exponential in costs. With this choice of $C(\cdot)$, the CoC's problem is to find a $\hat{\theta}$ that minimizes $\sup_{B \in \mathcal{B}} \ln\left(R_{\max}\left(B, \hat{\theta}\right)/R^*(B)\right) = \sup_{B \in \mathcal{B}} \ln A(B, \hat{\theta})$. The CoC will therefore propose the optimally adaptive estimator $\hat{\theta}^*$, which yields $\sup_{B \in \mathcal{B}} \Delta_B(\hat{\theta}^*) \propto \ln A^*(\mathcal{B})$. When $\tilde{C}(\ln A^*(\mathcal{B})) > F$, the CoC balks at the cost of implementing $\hat{\theta}^*$ and consensus fails.

A.3 Discussion

Taking the committees to represent different camps of researchers, the model suggests adaptive estimation can help to forge consensus between researchers with varying beliefs about the suitability of different econometric models. The prospects for achieving consensus are governed by the loss of efficiency under adaptation. When $A^*(\mathcal{B})$ is small, consensus is likely, as the adaptive estimator will yield maximal risk similar to each camp's perceived B-minimax risk. When $A^*(\mathcal{B})$ is large, however, consensus is unlikely to emerge, as the optimally adaptive estimator will be perceived as excessively risky by camps with extreme beliefs.

Appendix B Details and proofs

B.1 Details for Theorem 4.1 and extensions

We provide details and formal results for the results in Section 4.2 giving B-minimax and optimally adaptive estimators. We first provide a general theorem characterizing minimax estimators in a setting that includes our main example. We then specialize this result to derive the formula for the B-minimax estimator and optimally adaptive estimator for our main example given in Section 4.2, using a weighted loss function and Lemma 4.1 to obtain the optimally adaptive estimator. This proves Theorem 4.1.

We consider a slightly more general setting with p misspecified estimates, leading to a $p \times 1$ vector Y_O:

$$Y = \begin{pmatrix} Y_U \\ {\scriptstyle 1\times 1} \\ Y_O \\ {\scriptstyle p\times 1} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta \\ {\scriptstyle 1\times 1} \\ b \\ {\scriptstyle p\times 1} \end{pmatrix}, \Sigma \right), \quad \Sigma = \begin{pmatrix} \Sigma_U & \Sigma_{UO} \\ {\scriptstyle 1\times 1} & {\scriptstyle 1\times p} \\ \Sigma'_{UO} & \Sigma_O \\ {\scriptstyle p\times 1} & {\scriptstyle p\times p} \end{pmatrix}. \tag{10}$$

In our main example, $p = 1$ and $\rho = \Sigma_{UO}/\sqrt{\Sigma_U \Sigma_O}$. We are interested in the minimax risk of an estimator $\delta : \mathbb{R}^{p+1} \to \mathbb{R}$ under the loss function $L(\theta, b, d)$, which may incorporate a scaling to turn the minimax problem into a problem of finding an optimally adaptive estimator, following Lemma 4.1. We assume that the loss function satisfies the invariance condition

$$L(\theta + t, b, d + t) = L(\theta, b, d) \quad \text{all } t \in \mathbb{R}. \tag{11}$$

We consider minimax estimation over a parameter space $\mathbb{R} \times \mathcal{C}$:

$$\inf_{\delta} \sup_{\theta \in \mathbb{R}, b \in \mathcal{C}} R(\theta, b, \delta). \tag{12}$$

Theorem B.1. *Suppose that the loss function $L(\theta, b, d)$ is convex in d and that (11) holds.*

Then the minimax risk (12) is given by

$$\begin{aligned} &\inf_{\bar{\delta}} \sup_{b \in \mathcal{C}} E_{0,b}[\tilde{L}(b, \bar{\delta}(Y_O) - \Sigma_{UO}\Sigma_O^{-1} b)] \\ &= \sup_{\pi \text{ supported on } \mathcal{C}} \inf_{\bar{\delta}} \int E_{0,b}[\tilde{L}(b, \bar{\delta}(Y_O) - \Sigma_{UO}\Sigma_O^{-1} b)]\, d\pi(b) \end{aligned} \tag{13}$$

where $\tilde{L}(b,t) = EL(0,b,t+V)$ *with* $V \sim N(0, \Sigma_U - \Sigma_{UO}\Sigma_O^{-1}\Sigma_{UO}')$. *Furthermore, the minimax problem (12) has at least one solution, and any solution* δ^* *takes the form*

$$\delta^*(Y_U, Y_O) = Y_U - \Sigma_{UO}\Sigma_O^{-1} Y_O + \bar{\delta}^*(Y_O)$$

where $\bar{\delta}^*$ *achieves the infimum in (13).*

Proof. The minimax problem (12) is invariant (in the sense of pp. 159-161 of Lehmann and Casella (1998)) to the transformations $(\theta, b) \mapsto (\theta + t, b)$ and the associated transformation of the data $(Y_U, Y_O) \mapsto (Y_U + t, Y_O)$, where t varies over $\mathbb{R}$. Equivariant estimators for this group of transformations are those that satisfy $\delta(y_U + t, y_O) = \delta(y_U, y_O) + t$, which is equivalent to imposing that the estimator takes the form $\delta(y_U, y_O) = \delta(0, y_O) + y_U$. The risk of such an estimator does not depend on θ and is given by

$$R(\theta, b, \delta) = R(0, b, \delta) = E_{0,b}\left[L(0, b, \delta(0, Y_O) + Y_U)\right].$$

Using the decomposition $Y_U - \theta = \Sigma_{UO}\Sigma^{-1}(Y_O - b) + V$ where $V \sim N(0, \Sigma_U - \Sigma_{UO}\Sigma_O^{-1}\Sigma_{UO}')$ is independent of Y_O, the above display is equal to

$$E_{0,b}\left[L(0, b, \delta(0, Y_O) + \Sigma_{UO}\Sigma_O^{-1}(Y_O - b) + V)\right] = E_{0,b}\tilde{L}(b, \delta(0, Y_O) + \Sigma_{UO}\Sigma_O^{-1}(Y_O - b)).$$

Letting $\bar{\delta}(Y_O) = \delta(0, Y_O) + \Sigma_{UO}\Sigma_O^{-1} Y_O$, the above display is equal to $E_{0,b}[\tilde{L}(b, \bar{\delta}(Y_O) - \Sigma_{UO}\Sigma_O^{-1} b)]$. Thus, if an estimator $\bar{\delta}^*$ achieves the infimum in (13), the corresponding estimator $\delta(Y_U, Y_O) = \delta(0, Y_O) + Y_U = \bar{\delta}^*(Y_O) - \Sigma_{UO}\Sigma_O^{-1} Y_O + Y_U$ will be minimax among equivariant estimators for (12). It will then follow from the Hunt-Stein Theorem (Lehmann and Casella, 1998, Theorem 9.2) that this minimax equivariant estimator is minimax among all estimators, that any other minimax estimator takes this form and that the minimax risk is given by the first line of (13).

It remains to show that the infimum in the first line of (13) is achieved, and that the equality claimed in (13) holds. The equality in (13) follows from the minimax theorem, as stated in Theorem A.5 in Johnstone (2019) (note that $d \mapsto \tilde{L}(b, d - \Sigma_{UO}\Sigma_O^{-1}b)$ is convex since it is an integral of the convex functions $d \mapsto L(0, b, d - \Sigma_{UO}\Sigma_O^{-1}b + v)$ over the index v). The existence of an estimator $\bar{\delta}^*$ that achieves the infimum in the first line of (13) follows by noting that the set of decision rules (allowing for randomized decision rules) is compact in the topology defined on p. 405 of Johnstone (2019), and the risk $E_{0,b}[\tilde{L}(b, \bar{\delta}(Y_O) - \Sigma_{UO}\Sigma_O^{-1}b)]$ is continuous in $\bar{\delta}$ under this topology. As noted immediately after Theorem A.1 in Johnstone (2019), this implies that $\bar{\delta} \mapsto \sup_b E_{0,b}[\tilde{L}(b, \bar{\delta}(Y_O) - \Sigma_{UO}\Sigma_O^{-1}b)]$ is a lower semicontinuous function on the compact set of possibly randomized decision rules under this topology, which means that there exists a decision rule that achieves the minimum. From this possibly randomized decision rule, we can construct a nonrandomized decision rule that achieves the minimum by constructing a nonrandomized decision rule with uniformly smaller risk by averaging, following Johnstone (2019, p. 404). □

We now prove Theorem 4.1 by specializing this result. Note that Σ_U and Σ_O correspond to σ_U^2 and σ_O^2 in the main text respectively, and that ρ in the main text is given by $\Sigma_{UO}/\sqrt{\Sigma_U\Sigma_O}$.

First, we derive the minimax estimator and minimax risk in (12) when $L(\theta, b, d) = (\theta - d)^2$ and $\mathcal{C} = [-B, B]$. We have $\tilde{L}(b, t) = E(t + V)^2 = t^2 + \Sigma_U - \Sigma_{UO}^2/\Sigma_O$. Thus, (13) becomes

$$\begin{aligned}
&\inf_{\bar{\delta}} \sup_{b \in [-B,B]} E_{0,b}\left[\left(\bar{\delta}(Y_O) - \frac{\Sigma_{UO}}{\Sigma_O} b\right)^2\right] + \Sigma_U - \frac{\Sigma_{UO}^2}{\Sigma_O} \\
&= \inf_{\bar{\delta}} \sup_{b \in [-B,B]} \frac{\Sigma_{UO}^2}{\Sigma_O} E_{0,b}\left[\left(\frac{\sqrt{\Sigma_O}}{\Sigma_{UO}} \bar{\delta}(Y_O) - \frac{b}{\sqrt{\Sigma_O}}\right)^2\right] + \Sigma_U - \frac{\Sigma_{UO}^2}{\Sigma_O}.
\end{aligned}$$

This is equivalent to observing $T_O = Y_O/\sqrt{\Sigma_O} \sim N(t, 1)$ and finding the minimax estimator of t under the constraint $|t| \leq B/\sqrt{\Sigma_O}$. Letting $\delta^{\mathrm{BNM}}(T_O; B/\sqrt{\Sigma_O})$ denote the solution to this minimax problem and letting $r^{\mathrm{BNM}}(B/\sqrt{\Sigma_O})$ denote the value of this minimax problem, the optimal $\bar{\delta}$ in the above display satisfies $\frac{\sqrt{\Sigma_O}}{\Sigma_{UO}} \bar{\delta}(Y_O) = \delta^{\mathrm{BNM}}(Y_O/\sqrt{\Sigma_O}; B/\sqrt{\Sigma_O})$, which gives the value of the above display as

$$\frac{\Sigma_{UO}^2}{\Sigma_O} r^{\mathrm{BNM}}(B/\sqrt{\Sigma_O}) + \Sigma_U - \frac{\Sigma_{UO}^2}{\Sigma_O} \tag{14}$$

and the B-minimax estimator as

$$\frac{\Sigma_{UO}}{\sqrt{\Sigma_O}}\delta^{\text{BNM}}(Y_O/\sqrt{\Sigma_O}; B/\sqrt{\Sigma_O}) + Y_U - \frac{\Sigma_{UO}}{\Sigma_O}Y_O. \tag{15}$$

Substituting $T_O = Y_O/\sqrt{\Sigma_O}$ and the notation $\rho = \Sigma_{UO}/\sqrt{\Sigma_U\Sigma_O}$, $\sigma_U^2 = \Sigma_U$ and $\sigma_O^2 = \Sigma_O$ used in the main text gives (4) and (5). This proves part (i) of Theorem 4.1.

To find the optimally adaptive estimator and loss of efficiency under adaptation in our main example, we apply Lemma 4.1 with $\omega(\theta, b) = R^*(|b|)^{-1}$, with $R^*(B)$ given by (14). This leads to the minimax problem (12) with $\mathcal{C} = \mathbb{R}$ and $L(\theta, b, d) = R^*(|b|)^{-1}(\theta - d)^2$. The function $\tilde{L}$ in Theorem B.1 is then given by $\tilde{L}(b, t) = ER^*(|b|)^{-1}(t + V)^2 = R^*(|b|)^{-1}(t^2 + \Sigma_U - \Sigma_{UO}^2/\Sigma_O)$, which gives (13) as

$$\inf_{\bar{\delta}}\sup_{b\in\mathbb{R}}\frac{E_{0,b}\left[\left(\bar{\delta}(Y_O) - \frac{\Sigma_{UO}}{\Sigma_O}b\right)^2\right] + \Sigma_U - \frac{\Sigma_{UO}^2}{\Sigma_O}}{\frac{\Sigma_{UO}^2}{\Sigma_O}r^{\text{BNM}}(|b|/\sqrt{\Sigma_O}) + \Sigma_U - \frac{\Sigma_{UO}^2}{\Sigma_O}} = \inf_{\bar{\delta}}\sup_{b\in\mathbb{R}}\frac{E_{0,b}\left[\left(\frac{\sqrt{\Sigma_O}}{\Sigma_{UO}}\bar{\delta}(Y_O) - \frac{b}{\sqrt{\Sigma_O}}\right)^2\right] + \rho^{-2} - 1}{r^{\text{BNM}}(|b|/\sqrt{\Sigma_O}) + \rho^{-2} - 1}.$$

This proves part (iii) of Theorem 4.1. The above display is minimized by $\bar{\delta}$ satisfying $\frac{\sqrt{\Sigma_O}}{\Sigma_{UO}}\bar{\delta}(Y_O) = \delta^*(Y_O/\sqrt{\Sigma_O}; \rho)$ where $\delta^*(T; \rho)$ minimizes (6) in the main text. By Theorem B.1, the optimally adaptive estimator is given by

$$\frac{\Sigma_{UO}}{\sqrt{\Sigma_O}}\delta^*(Y_O/\sqrt{\Sigma}; \rho) + Y_U - \frac{\Sigma_{UO}}{\Sigma_O}Y_O = \rho\sqrt{\Sigma_U}\delta^*(T_O; \rho) + Y_U - \rho\sqrt{\Sigma_U}T_O. \tag{16}$$

This proves the part (ii) of Theorem 4.1.

B.2 Lasso interpretation of soft thresholding

To connect the soft thresholding estimator to lasso, consider a dataset with two observations comprised of the realizations of Y_U and Y_R, and a linear model relating these estimates to a constant and an indicator for whether the observation is from the restricted specification. Letting $y_1 = Y_U$, $d_1 = 0$, $y_2 = Y_R$, and $d_2 = 1$, the model can be written

$$y_i = \beta + d_i\gamma + u_i,$$

where $\beta = \theta$, $\gamma = b$. Now consider an ℓ_1-penalized GLS regression estimator

$$(\hat{\beta}'_{lasso,\lambda}, \hat{\gamma}_{lasso,\lambda}) = \arg\min_{\beta,\gamma} \frac{1}{2}\|\tilde{y} - \tilde{X}\beta - \tilde{z}\gamma\|_2^2 + \lambda|\gamma|,$$

where $\tilde{y}$, $\tilde{z}$, and $\tilde{X}$ are transformed so that the observations are orthogonalized and standardized.

Theorem B.2. *Suppose that the lasso penalty λ is set to equal to the adaptive soft threshold (divided by σ_O). Then the lasso regression coefficient estimator*

$$\hat{\beta}_{lasso,\lambda} = Y_{GMM} + \rho\sigma_U \delta_{S,\lambda\sigma_O}(T_O).$$

is the same as the soft thresholding nearly adaptive estimator.

Proof. We first prove a general representation of the lasso regression coefficient estimator as a soft-thresholding estimator, and then we specialize the result to our setting. Consider a penalized regression estimator

$$(\hat{\beta}'_{\text{Pen},\lambda}, \hat{\gamma}_{\text{Pen},\lambda}) = \arg\min_{\beta,\gamma} \frac{1}{2}\|y - X\beta - z\gamma\|_2^2 + \lambda\,\text{Pen}(\gamma) \tag{17}$$

where y and Z are $n \times 1$ vectors and X is a $n \times k$ matrix. We use $P_X = X(X'X)^{-1}X'$ and $M_X = I - P_X$ to denote the projection onto the column space of X and onto its orthogonal complement. We are interested in the scalar parameter $\ell'\beta$ for some known vector ℓ and wish to compare the estimator $\ell'\hat{\beta}_{\text{Pen},\lambda}$ to estimators that are optimally adaptive or constrained optimally adaptive for $\ell'\beta$ under a restriction on the bias of the short regression estimator $\ell'\hat{\beta}_{\text{short}}$ where $\hat{\beta}_{\text{short}} = (X'X)^{-1}X'y$.

Note that standard regression algebra immediately implies that $\hat{\beta}_{\text{Pen},\lambda}$ can be obtained by regressing $y - z\hat{\gamma}_{\text{Pen},\lambda}$ on X, which gives

$$\ell'\hat{\beta}_{\text{Pen},\lambda} = \ell'(X'X)^{-1}X'(y - z\hat{\gamma}_{\text{Pen},\lambda}) = \ell'\hat{\beta}_{\text{short}} - \ell'(X'X)^{-1}X'z\hat{\gamma}_{\text{Pen},\lambda}. \tag{18}$$

To derive $\hat{\gamma}_{\text{Pen},\lambda}$, note that the objective in (17) can be written as

$$\frac{1}{2}\|M_X y - M_X z\gamma\|_2^2 + \frac{1}{2}\|P_X(y - z\gamma) - X\beta\|_2^2 + \lambda\,\text{Pen}(\gamma).$$

Since the second term can be set to zero for any value of γ by taking $\beta = (X'X)^{-1}X'(y - z\gamma)$, and β does not show up in the remaining terms, it follows that this term can be ignored when optimizing $\hat{\gamma}_{\text{Pen},\lambda}$. Thus, $\hat{\gamma}_{\text{Pen},\lambda}$ minimizes

$$\frac{1}{2}\|M_X y - M_X z\gamma\|_2^2 + \lambda \,\text{Pen}(\gamma).$$

Consider the lasso case where $\text{Pen}(\gamma) = |\gamma|$. Taking FOCs gives

$$\begin{aligned} &- z'M_X(y - z\gamma) + \lambda \,\text{sign}(\gamma) = 0 \\ &\iff \gamma = \frac{z'M_X y}{z'M_X z} - \frac{\lambda}{z'M_X z}\,\text{sign}(\gamma) = \hat{\gamma}_{\text{long}} - \frac{\lambda}{z'M_X z}\,\text{sign}(\gamma) \end{aligned}$$

where $\text{sign}(\gamma)$ is the set-valued function equal to the sign of γ when γ is nonzero, and equal to $[-1, 1]$ when $\gamma = 0$. There are three cases to consider. First, if $\hat{\gamma}_{\text{long}} > \lambda/z'M_X z$, then $\text{sign}(\gamma) = 1$ so that $\gamma = \hat{\gamma}_{\text{long}} - \lambda/z'M_X z$. Second, if $\hat{\gamma}_{\text{long}} < -\lambda/z'M_X z$, then $\text{sign}(\gamma) = -1$ so that $\gamma = \hat{\gamma}_{\text{long}} + \lambda/z'M_X z$. Finally, if $\hat{\gamma}_{\text{long}} \in [-\lambda/z'M_X z, \lambda/z'M_X z]$, then we will run into a contradiction if $\gamma \neq 0$: $\gamma > 0$ would imply $\text{sign}(\gamma) = 1$ which would give $\gamma = \hat{\gamma}_{\text{long}} - \lambda/z'M_X z \leq 0$ and $\gamma < 0$ would imply $\text{sign}(\gamma) = -1$ which would give $\gamma = \hat{\gamma}_{\text{long}} + \lambda/z'M_X z \geq 0$. Thus, if $\hat{\gamma}_{\text{long}} \in [-\lambda/z'M_X z, \lambda/z'M_X z]$, we must have $\gamma = 0$. It follows that the solution to the optimization problem is given by

$$\hat{\gamma}_{\text{Pen},\gamma} = \begin{cases} 0 & \text{when } |\hat{\gamma}_{\text{long}}| \leq |\lambda/z'M_X z| \\ \hat{\gamma}_{\text{long}} - \lambda/z'M_X z & \text{when } \hat{\gamma}_{\text{long}} > \lambda/z'M_X z \\ \hat{\gamma}_{\text{long}} + \lambda/z'M_X z & \text{when } \hat{\gamma}_{\text{long}} < \lambda/z'M_X z \end{cases}$$

This is the soft threshold estimator $\delta_{S,\lambda/z'M_X z}(\hat{\gamma}_{\text{long}})$ with cutoff $\lambda/z'M_X z$. Plugging this into (18) gives the penalized regression estimate for our parameter of interest as

$$\ell'\hat{\beta}_{\text{Pen},\lambda} = \ell'\hat{\beta}_{\text{short}} - \ell'(X'X)^{-1}X'z \cdot \delta_{S,\lambda/z'M_X z}(\hat{\gamma}_{\text{long}})$$

Now apply the GLS transformation to the data as the follows

$$\tilde{y} = \begin{pmatrix} Y_{GMM}/\sigma_{R,GMM} \\ T_O \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_{R,GMM}} & 0 \\ 0 & \frac{1}{\sigma_O} \end{pmatrix} \begin{pmatrix} 1 + \rho\frac{\sigma_U}{\sigma_O} & -\rho\frac{\sigma_U}{\sigma_O} \\ -1 & 1 \end{pmatrix} \begin{pmatrix} Y_U \\ Y_R \end{pmatrix},$$

$$\tilde{X} = \begin{pmatrix} \frac{1}{\sigma_{R,GMM}} & 0 \\ 0 & \frac{1}{\sigma_O} \end{pmatrix} \begin{pmatrix} 1+\rho\frac{\sigma_U}{\sigma_O} & -\rho\frac{\sigma_U}{\sigma_O} \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_{R,GMM}} \\ 0 \end{pmatrix}$$

and

$$\tilde{z} = \begin{pmatrix} \frac{1}{\sigma_{R,GMM}} & 0 \\ 0 & \frac{1}{\sigma_O} \end{pmatrix} \begin{pmatrix} 1+\rho\frac{\sigma_U}{\sigma_O} & -\rho\frac{\sigma_U}{\sigma_O} \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -\frac{1}{\sigma_{R,GMM}} \cdot \rho\frac{\sigma_U}{\sigma_O} \\ \frac{1}{\sigma_O} \end{pmatrix}.$$

The least squares estimator of γ is the minimum variance unbiased estimate for $\gamma = b$, which is $\hat{\gamma}_{\text{long}} = Y_O$. The short regression estimator of β in the transformed model is $\hat{\beta}_{\text{short}} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} = Y_{GMM}$. Finally, $(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{z} = \sigma^2_{R,GMM} \cdot \frac{1}{\sigma_{R,GMM}} \cdot \frac{-1}{\sigma_{R,GMM}} \cdot \frac{\rho\sigma_U}{\sigma_O} = -\rho\frac{\sigma_U}{\sigma_O}$ and $\tilde{z}'M_{\tilde{X}}\tilde{z} = 1/\sigma^2_O$. Thus, the GLS lasso estimate is

$$Y_{GMM} + \rho\frac{\sigma_U}{\sigma_O}\delta_{S,\lambda\sigma^2_O}(Y_O).$$

Note that soft thresholding Y_O at $\lambda\sigma^2_O$ is equivalent to soft thresholding $T_O = Y_O/\sigma_O$ at $\lambda\sigma_O$ and multiplying by σ_O. Thus, we can also write the GLS lasso estimate as

$$Y_{GMM} + \rho\sigma_U\delta_{S,\lambda\sigma_O}(T_O).$$

This is the same as the soft thresholding nearly adaptive estimator, but with λ replaced by $\lambda \cdot \sigma_O$.

□

Online Appendix to "Adapting to Misspecification"

Timothy B. Armstrong, Patrick Kline and Liyang Sun

August 2024

Appendix C Additional details

C.1 Constrained adaptation

The constrained adaptive estimator solves the problem

$$A^*(\mathcal{B}; \overline{R}) = \inf_{\hat{\theta}} \sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \hat{\theta})}{R^*(B)} \quad \text{s.t.} \quad \sup_{B \in \mathcal{B}} R_{\max}(B, \hat{\theta}) \leq \overline{R}. \tag{19}$$

We can rewrite this formulation as a weighted minimax problem similar to the one in Section 4.1 by setting $t = \overline{R}/A^*(\mathcal{B}; \overline{R})$ and considering the problem

$$\inf_{\hat{\theta}} \sup_{B \in \mathcal{B}} \max \left\{ \frac{R_{\max}(B, \hat{\theta})}{R^*(B)}, \frac{R_{\max}(B, \hat{\theta})}{t} \right\} = \inf_{\hat{\theta}} \sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \hat{\theta})}{\min\{R^*(B), t\}}. \tag{20}$$

Indeed, any solution to (19) must also be a solution to (20) with $t = \overline{R}/A^*(\mathcal{B}; \overline{R})$, since any decision function achieving a strictly better value of (20) would satisfy the constraint in (19) and achieve a strictly better value of the objective in (19). Conversely, letting $\tilde{A}^*(t)$ be the value of (20), any solution to (20) will achieve the same value of the objective (19) and will satisfy the constraint for $\bar{R} = t \cdot \tilde{A}^*(t)$. In fact, this solution to (20) will also solve (19) for $\bar{R} = t \cdot \tilde{A}^*(t)$ so long as this value of $\bar{R}$ is large enough to allow some scope for adaptation.

Arguing as in Section 4.1, we can write the optimization problem (20) as

$$\inf_{\hat{\theta}} \sup_{(\theta,b) \in \cup_{B' \in \mathcal{B}} \mathcal{C}_{B'}} \tilde{\omega}(\theta, b, t) R(\theta, b, \hat{\theta}), \tag{21}$$

$$\text{where } \tilde{\omega}(\theta, b, t) = \left(\inf_{B \in \mathcal{B} \text{ s.t. } (\theta,b) \in \mathcal{C}_B} \min\{R_{\max}(B), t\} \right)^{-1} = \max\{\omega(\theta, b), 1/t\}$$

and $\omega(\theta, b)$ is given in Lemma 4.1 in Section 4.1. Thus, we can solve (20) by solving for the minimax estimator under the loss function $(\theta, b, d) \mapsto \tilde{\omega}(\theta, b, t)L(\theta, b, d)$. Letting $A^*(t)$ be the optimized objective function, we can then solve (19) by finding a t such that $\bar{R} = t \cdot A^*(t)$.

We summarize these results in the following lemma, which is proved in Section C.1.1 of the appendix.

Lemma C.1. *Any solution to (19) is also a solution to (21) with $t = \overline{R}/A^*(\mathcal{B}; \overline{R})$. Conversely, let $\tilde{A}^*(t)$ denote the value of (21) and let $\tilde{R}(t) = \tilde{A}^*(t) \cdot t$. If $\tilde{R}(t) > \inf_{\hat{\theta}} \sup_{B \in \mathcal{B}} R_{\max}(B, \hat{\theta})$ and $\inf_{B \in \mathcal{B}} R^*(B) > 0$, then $A^*(\mathcal{B}; \tilde{R}(t)) = \tilde{A}^*(t)$ and any solution to (21) is also a solution to (19) with $\bar{R} = \tilde{R}(t)$.*

C.1.1 Details for constrained adaptation

We provide proof for Lemma C.1, which shows the constrained adaption problem is equivalent to the weighted minimax problem with a particular set of weights. The first statement is immediate from the arguments proceeding the statement of the lemma in Section 4.4. For the second statement, let $\bar{\delta}$ be a decision rule with $\sup_{B \in \mathcal{B}} R_{\max}(B, \bar{\delta}) < \tilde{R}(t)$. Such a decision rule exists and satisfies $\sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \bar{\delta})}{R^*(B)} < \infty$ by the assumptions of the lemma. Let δ_t^* be a solution to (20).

Suppose, to get a contradiction, that a decision δ' satisfies the constraint in (19) with $\bar{R} = \tilde{R}(t)$ and achieves a strictly better value of the objective than $\tilde{A}^*(t)$. For $\lambda \in (0, 1)$, let δ'_λ be the randomized decision rule that places probability λ on $\bar{\delta}$ and probability $1 - \lambda$ on δ', independently of the data Y. Note that $R_{\max}(B, \delta'_\lambda) = \sup_{(\theta,b) \in \mathcal{C}_B} R(\theta, b, \delta'_\lambda) = \sup_{(\theta,b) \in \mathcal{C}_B} \left[\lambda R(\theta, b, \bar{\delta}) + (1 - \lambda)R(\theta, b, \delta')\right] \leq \sup_{(\theta,b) \in \mathcal{C}_B} \lambda R(\theta, b, \bar{\delta}) + \sup_{(\theta,b) \in \mathcal{C}_B} (1-\lambda)R(\theta, b, \delta') = \lambda R_{\max}(B, \bar{\delta}) + (1 - \lambda)R_{\max}(B, \delta')$ so that, for $\lambda \in (0, 1)$,

$$\sup_{B \in \mathcal{B}} R_{\max}(B, \delta_\lambda) \leq \lambda \sup_{B \in \mathcal{B}} R_{\max}(B, \bar{\delta}) + (1 - \lambda) \sup_{B \in \mathcal{B}} R_{\max}(B, \delta') < \tilde{R}(t) = \tilde{A}^*(t) \cdot t$$

and

$$\sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \delta_\lambda)}{R^*(B)} \leq \lambda \sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \bar{\delta})}{R^*(B)} + (1 - \lambda) \sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \delta')}{R^*(B)}.$$

Since $\sup_{B \in \mathcal{B}} \frac{R_{\max}(B, \bar{\delta})}{R^*(B)}$ is finite and $\frac{\sup_{B \in \mathcal{B}} R_{\max}(B, \delta')}{R^*(B)} < \tilde{A}^*(t)$, the above display is strictly less than $\tilde{A}^*(t)$ for small enough λ. Thus, for small enough λ, the objective function in (21)

evaluated at the decision function δ_λ evaluates to

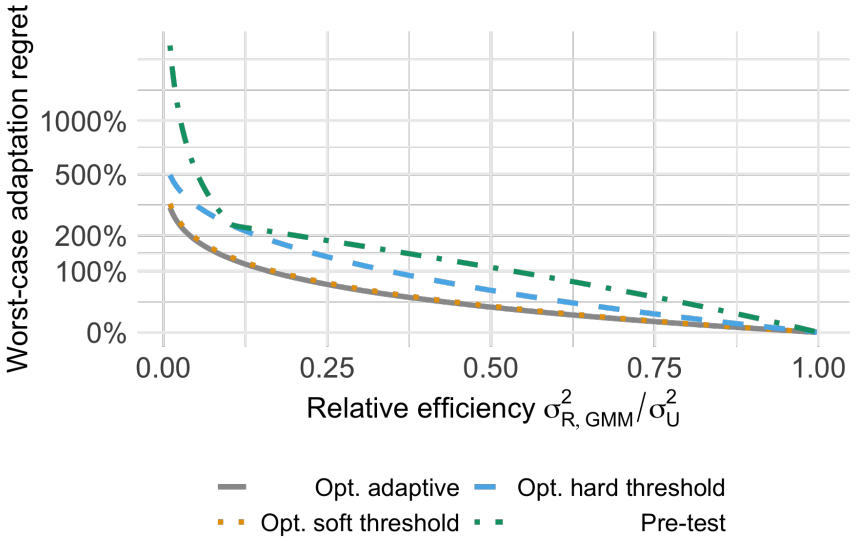
$$\max\left\{\sup_{B\in\mathcal{B}}\frac{R_{\max}(B,\delta_\lambda)}{R^*(B)},\sup_{B\in\mathcal{B}}\frac{R_{\max}(B,\delta_\lambda)}{t}\right\}<\max\left\{\tilde{A}^*(t),\tilde{R}(t)/t\right\}=\tilde{A}^*(t),$$

a contradiction.

C.2 Numerical results on estimators as a function of $1-\rho^2$

In practice, it is common to use a fixed threshold of 1.96, which corresponds to a pre-test rule that switches between the unrestricted estimator and the GMM estimator based on the result of the specification test. Doing so leads to high level of worst-case adaptation regret especially when ρ^2 is close to one as shown in Figure A1. To minimize the worst-case adaptation regret, the adaptive hard-threshold estimator needs to use a threshold that would increase to infinity as ρ^2 gets closer to one.

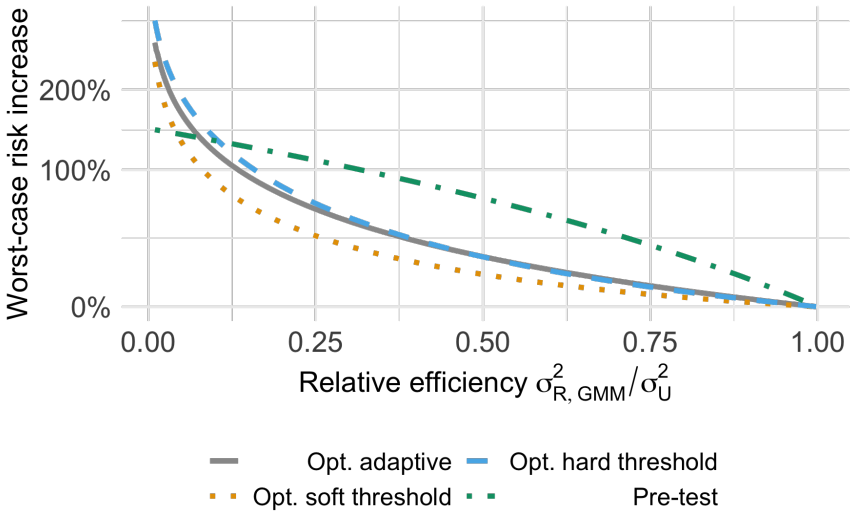
Figure A1: Worst case adaptation regret as function of relative efficiency



Notes: Vertical axis plots $(A_{\max}(\mathcal{B},\hat{\theta})-1)\times 100$ on $\log_{10}$ scale.

A pre-test estimator utilizing a fixed threshold at 1.96 realizes its worst-case risk when the scaled bias $\tilde{b}$ is itself near the 1.96 threshold. As shown in Figure A2, the pre-test estimator tends to exhibit substantially greater worst-case risk than the class of adaptive estimators for most values of ρ^2. As discussed in Section C.3 below, adaptive estimators have large worst-case risk when ρ^2 is close to one. The pre-test estimator has lower worst-case risk in these cases, due to the fixed threshold at 1.96.

Figure A2: Worst case risk increase relative to Y_U



Notes: Vertical axis plots $(R_{\max}(\infty, \hat{\theta}) - \sigma_U)/\sigma_U \times 100$ on $\log_{10}$ scale.

C.3 Asymptotics as $|\rho| \to 1$

This section considers the behavior of the worst-case adaptation regret as $|\rho| \to 1$ for the optimally adaptive estimator as well as for the hard and soft-thresholding estimators. Recall that $1 - \rho^2$ is equal to $\sigma^2_{R,GMM}/\sigma^2_U$, so that $|\rho| \to 1$ corresponds to the case where $\sigma^2_{R,GMM}/\sigma^2_U \to 0$. It will be convenient to phrase our results in terms of $\rho^{-2} - 1 = (1 - \rho^2)/\rho^2 = (1 + o(1)) \cdot \sigma^2_{R,GMM}/\sigma^2_U$ as $|\rho| \to 1$.

Let $A(\delta, \rho)$ denote the worst-case adaptation regret of the estimator given by (4) under the given value of ρ, so that $A(\delta, \rho)$ returns the value of (6) with $\tilde{\delta} = \delta$. We use $A^*(\rho) = \inf_\delta A(\delta, \rho)$ (where the infimum is over all estimators) to denote the loss of efficiency under adaptation for the given value of ρ. Likewise, we denote by $A_S(\lambda, \rho) = A(\delta_{S,\lambda}, \rho)$ and $A_H(\lambda, \rho) = A(\delta_{H,\lambda}, \rho)$ the worst-case adaptation regret for soft and hard-thresholding respectively with threshold λ, where $\delta_{S,\lambda}$ are $\delta_{H,\lambda}$ are defined in Section 4.3. Finally, we use $A^*_S(\rho) = \inf_\lambda A_S(\lambda, \rho)$ and $A^*_H(\rho) = \inf_\lambda A_H(\lambda, \rho)$ to denote the minimum worst-case adaptation regret for soft and hard-thresholding respectively.

The following theorem characterizes the behavior of $A^*(\rho)$, $A^*_S(\rho)$ and $A^*_H(\rho)$ as $|\rho| \to 1$.

Theorem C.1. *We have*

$$\lim_{|\rho| \uparrow 1} \frac{A^*(\rho)}{2 \log(\rho^{-2} - 1)^{-1}} = \lim_{|\rho| \uparrow 1} \frac{A^*_S(\rho)}{2 \log(\rho^{-2} - 1)^{-1}} = \lim_{|\rho| \uparrow 1} \frac{A^*_H(\rho)}{2 \log(\rho^{-2} - 1)^{-1}} = 1.$$

In the remainder of this section, we prove Theorem C.1. We split the proof into upper bounds (Section C.3.1) and lower bounds (Section C.3.2). The lower bounds in Section C.3.2 are essentially immediate from results in Bickel (1983) for adapting to $B \in \mathcal{B} = \{0, \infty\}$, whereas the upper bounds in Section C.3.1 involve new arguments to deal with intermediate values of B.

C.3.1 Upper bounds

In this section, we show that $A_S^*(\rho) \leq (1 + o(1))2\log(\rho^{-2} - 1)^{-1}$ and $A_H^*(\rho) \leq (1 + o(1))2\log(\rho^{-2} - 1)^{-1}$. Since $A^*(\rho)$ is bounded from above by both $A_S^*(\rho)$ and $A_H^*(\rho)$, this also implies $A^*(\rho) \leq (1 + o(1))2\log(\rho^{-2} - 1)^{-1}$.

Let $r_S(\lambda, t) = E_{T\sim N(\mu,1)}(\delta_{S,\lambda}(T) - \mu)^2$ and $r_S(\lambda, t) = E_{T\sim N(\mu,1)}(\delta_{H,\lambda}(T) - \mu)^2$ denote the risk of soft and hard-thresholding. Then

$$A_S(\lambda, \rho) = \sup_{\mu\in\mathbb{R}} \frac{r_S(\lambda, \mu) + \rho^{-2} - 1}{r^{\mathrm{BNM}}(|\mu|) + \rho^{-2} - 1}$$

and similarly for $A_H(\lambda, \rho)$. We use the following upper bound for $r_H(\lambda, \mu)$ and $r_S(\lambda, \mu)$, which follows immediately from results given in Johnstone (2019).

Lemma C.2. *There exists a constant C such that, for $\lambda > C$, both $r_S(\lambda, \mu)$ and $r_H(\lambda, \mu)$ are bounded from above by $\bar{r}(\lambda, \mu)$ where*

$$\bar{r}(\lambda, \mu) = \begin{cases} \min\left\{\lambda \exp\left(-\lambda^2/2\right) + 1.2\mu^2, 1 + \mu^2\right\} & |\mu| \leq \lambda \\ 1 + \lambda^2 & |\mu| > \lambda. \end{cases}$$

Proof. The bound for $r_H(\lambda, \mu)$ follows from Lemma 8.5 in Johnstone (2019) along with the bound $r_H(\lambda, 0) \leq \frac{2+\varepsilon}{\sqrt{2\pi}}\lambda \exp\left(-\lambda^2/2\right)$ which holds for any $\varepsilon > 0$ for λ large enough by (8.15) in Johnstone (2019). The bound for $r_L(\lambda, \mu)$ follows from Lemma 8.3 and (8.7) in Johnstone (2019). □

Let $\tilde{\lambda}_\rho = \sqrt{2\log(\rho^{-2} - 1)^{-1}}$. By Lemma C.2, $A_S^*(\rho)$ and $A_H^*(\rho)$ are, for $(\rho^{-2} - 1)^{-1}$ large enough, bounded from above by the supremum over μ of

$$\frac{\bar{r}(\tilde{\lambda}_\rho, \mu) + \rho^{-2} - 1}{r^{\mathrm{BNM}}(|\mu|) + \rho^{-2} - 1} \tag{22}$$

Let $c(\rho)$ be such that $c(\rho)/\tilde{\lambda}_\rho \to 0$ and $c(\rho) \to \infty$ as $|\rho| \uparrow 1$. We bound (22) separately for $|\mu| \le c(\rho)$ and for $|\mu| \ge c(\rho)$. For $|\mu| \le c(\rho)$, we use the bound $r^{\mathrm{BNM}}(|\mu|) \ge .8 \cdot \mu^2/(\mu^2+1)$ (Donoho, 1994), which gives an upper bound for (22) of

$$\frac{\bar{r}(\tilde{\lambda}_\rho, \mu) + \rho^{-2} - 1}{.8 \cdot \mu^2/(\mu^2+1) + \rho^{-2} - 1} \le \frac{\sqrt{2\log(\rho^{-2}-1)^{-1}} \cdot (\rho^{-2}-1) + 1.2\mu^2 + \rho^{-2} - 1}{.8 \cdot \mu^2/(\mu^2+1) + \rho^{-2} - 1}$$
$$\le \sqrt{2\log(\rho^{-2}-1)^{-1}} + (1.2/.8) \cdot (\mu^2+1) + 1 \le \sqrt{2\log(\rho^{-2}-1)^{-1}} + (1.2/.8) \cdot (c(\rho)^2+1) + 1.$$

As $|\rho| \uparrow 1$, this increases more slowly than $\log(\rho^{-2}-1)^{-1}$. For $|\mu| \ge c(\rho)$, we use the bound $r^{\mathrm{BNM}}(|\mu|) \ge r^{\mathrm{BNM}}(c(\rho))$ which gives an upper bound for (22) of

$$\frac{\bar{r}(\tilde{\lambda}_\rho, \mu) + \rho^{-2} - 1}{r^{\mathrm{BNM}}(|c(\rho)|) + \rho^{-2} - 1} \le \frac{\bar{r}(\tilde{\lambda}_\rho, \mu)}{r^{\mathrm{BNM}}(|c(\rho)|)} + 1 \le \frac{1 + \tilde{\lambda}_\rho^2}{r^{\mathrm{BNM}}(|c(\rho)|)} + 1.$$

As $|\rho| \uparrow 1$, $c(\rho) \to \infty$ and $r^{\mathrm{BNM}}(|c(\rho)|) \to 1$, so that the above display is equal to a $1 + o(1)$ term times $\tilde{\lambda}_\rho^2 = 2\log(\rho^{-2}-1)^{-1}$ as required.

C.3.2 Lower bounds

In this section, we show that $A^*(\rho) \ge (1+o(1))2\log(\rho^{-2}-1)^{-1}$. Since $A_S^*(\rho)$ and $A_H^*(\rho)$ are bounded from below by $A^*(\rho)$, this also implies $A_S^*(\rho) \ge (1+o(1))2\log(\rho^{-2}-1)^{-1}$ and $A_H^*(\rho) \ge (1+o(1))2\log(\rho^{-2}-1)^{-1}$.

Given an estimator $\delta(Y)$ of μ in the normal means problem $Y \sim N(\mu, 1)$, let $m(\delta) = E_{T\sim N(0,1)}\delta(Y)^2$ denote the risk at $\mu = 0$ and let $M(\delta) = \sup_{\mu\in\mathbb{R}} E_{T\sim N(\mu,1)}(\delta(Y)-\mu)^2$ denote worst-case risk. The following lemma is immediate from Bickel (1983, Theorem 4.1).

Lemma C.3 (Bickel 1983, Theorem 4.1)**.** *For $t \in (0,1]$, let δ_t be an estimator that satisfies $m(\delta_t) \le 1-t$. Then, as $t \uparrow 1$, $M(\delta_t) \ge (1+o(1)) \cdot 2\log(1-t)$.*

Using this result, we prove the following lemma, which gives a lower bound for the worst-case adaptation regret and the worst-case risk of any estimator achieving the upper bound in Section C.3.1. The required lower bound $A^*(\rho) \ge (1+o(1))2\log(\rho^{-2}-1)^{-1}$ follows from this result.

Lemma C.4. *For $\rho \in (-1,1)$, let $\delta_\rho : \mathbb{R} \to \mathbb{R}$ be an estimator of μ in the normal means problem $Y \sim N(\mu,1)$. Suppose that the worst-case adaptation regret $A(\delta_\rho, \rho)$ of the corre-*

sponding estimator (4) satisfies $A(\delta_\rho, \rho) \leq (1 + o(1))2\log(\rho^{-2} - 1)^{-1}$ *as* $|\rho| \to 1$. *Then the following results hold as* $|\rho| \to 1$.

i.) The worst-case risk of the corresponding estimator (4) is bounded from below by a $1 + o(1)$ *term times* $2\Sigma_U \log(\rho^{-2} - 1)^{-1}$

ii.) $A(\delta_\rho, \rho) \geq (1 + o(1)) \cdot 2\log(\rho^{-2} - 1)^{-1}$.

Proof. By the arguments Section B.1, the worst-case risk of the estimator (4) with $\delta = \delta_\rho$ is given by $\Sigma_U \cdot \left[\rho^2 \sup_\mu E_{T\sim N(\mu,1)}(\delta_\rho(T) - \mu)^2 + 1 - \rho^2\right]$. As $|\rho| \uparrow 1$, this is bounded from below by a $1 + o(1)$ term times $\Sigma_U \sup_\mu E_{T\sim N(\mu,1)}(\delta_\rho(T) - \mu)^2$. Similarly, $A(\delta_\rho, \rho)$ is bounded from below by a $1 + o(1)$ term times $\sup_\mu E_{T\sim N(\mu,1)}(\delta_\rho(T) - \mu)^2$ as $|\rho| \uparrow 1$. Thus, it suffices to show that $\sup_\mu E_{T\sim N(\mu,1)}(\delta_\rho(T) - \mu)^2 \geq (1 + o(1)) \cdot 2\log(\rho^{-2} - 1)^{-1}$.

To show this, note that it follows from plugging in $\tilde{b} = 0$ to the objective in (6) that, for any $\varepsilon > 0$, we have, for $|\rho|$ close enough to 1,

$$\frac{E_{T\sim N(0,1)}\delta_\rho(T)^2}{\rho^{-2} - 1} \leq A(\delta_\rho, \rho) \leq (2 + \varepsilon)\log(\rho^{-2} - 1)^{-1}.$$

Applying Lemma C.3 with $1 - t = (\rho^{-2} - 1) \cdot (2 + \varepsilon)\log(\rho^{-2} - 1)^{-1}$, it follows that

$$\begin{aligned}
&\sup_\mu E_{T\sim N(\mu,1)}(\delta_\rho(T) - \mu)^2 \geq (1 + o(1)) \cdot 2\log\left[(\rho^{-2} - 1) \cdot (2 + \varepsilon)\log(\rho^{-2} - 1)^{-1}\right] \\
&= (1 + o(1)) \cdot \left[2\log(\rho^{-2} - 1) + \log(2 + \varepsilon) + \log\log(\rho^{-2} - 1)^{-1}\right] = (1 + o(1)) \cdot 2\log(\rho^{-2} - 1)
\end{aligned}$$

as required.

□

Appendix D Computational details

In this section, we provide additional details on our computation of the adaptive estimator.

D.1 Computing minimax estimators

As shown in Sections 4.1 and 4.2, one can compute adaptive estimators by solving a weighted minimax problem which, in our setting, can be further simplified using invariance. To solve

these problems, we use the insight that the minimax estimator can be characterized as a Bayes estimator for a *least favorable prior*. We first give a brief review of this approach before going into details for our setting.

Consider the generic problem of computing a minimax decision over the parameter space $\mathcal{C}$ for a parameter ϑ under loss $\bar{L}(\vartheta, \delta)$. We use E_ϑ and P_ϑ to denote expectation under ϑ and the probability distribution of the data Y under ϑ. Letting π denote a *prior* distribution on $\mathcal{C}$, the *Bayes risk* of δ is given by

$$R_{\text{Bayes}}(\pi, \delta) = \int E_\vartheta \bar{L}(\vartheta, \delta(Y))\, d\pi(\vartheta) = \int \int \bar{L}(\vartheta, \delta(y))\, dP_\vartheta(y) d\pi(\vartheta).$$

The *Bayes decision*, which we will denote $\delta_\pi^{\text{Bayes}}$, optimizes $R_{\text{Bayes}}(\pi, \delta)$ over δ. It can be computed by optimizing expected loss under the posterior distribution for ϑ taking π as the prior. Under squared error loss, the Bayes decision is the posterior mean.

$R_{\text{Bayes}}(\pi, \delta)$ gives a lower bound for the worst-case risk of δ under $\mathcal{C}$ and $R_{\text{Bayes}}(\pi, \delta_\pi^{\text{Bayes}})$ gives a lower bound for the minimax risk. Under certain conditions, a *minimax theorem* applies, which tells us that this lower bound is in fact sharp. In this case, letting Γ denote the set of priors π supported on $\mathcal{C}$, the minimax risk over $\mathcal{C}$ is given by

$$\min_\delta \max_{\pi \in \Gamma} R_{\text{Bayes}}(\pi, \delta) = \max_{\pi \in \Gamma} \min_\delta R_{\text{Bayes}}(\pi, \delta) = \max_{\pi \in \Gamma} R_{\text{Bayes}}(\pi, \delta_\pi^{\text{Bayes}}).$$

The distribution π that solves this maximization problem is called the *least favorable prior*. When the minimax theorem applies, the Bayes decision for this prior is the minimax decision over $\mathcal{C}$.

The expression $R_{\text{Bayes}}(\pi, \delta_\pi^{\text{Bayes}})$ is convex as a function of π if the set of possible decision functions is sufficiently unrestricted and the set Γ is convex. While one may need to allow randomized decisions in general, the estimation problems we consider will be such that the Bayes decision is nonrandomized. Thus, we can use convex optimization software to compute the least favorable prior and minimax estimator so long as we have a way of approximating π with a finite dimensional object that retains the convex structure of the problem.

In our setting, we use invariance arguments to obtain the objective function (6), which is a minimax problem over the unknown parameter $\tilde{b} = b/\sigma_O$ (the noncentrality parameter of the overidentification statistic T_O). We solve (6), as well as the bounded normal mean problem used to obtain the scaling in (6), by solving for a least favorable prior over $\tilde{b}$ using

a finite dimensional approximation $\pi(\tilde{b}_1), \ldots, \pi(\tilde{b}_J)$ to the prior over a grid of J values of $\tilde{b}$. The least favorable prior for (θ, b) is then given by a flat (improper) prior for θ along with the corresponding prior for $\tilde{b} = b/\sigma_O$, with the flat prior for θ following from invariance. We now discuss the details of this approximation.

D.2 Discrete approximation to estimators and risk function

Operationally, discretizing the support of the random variable $T \in \mathcal{T}$ into K points, finding an estimator $\delta(T)$ is equivalent to finding a "policy" function $\delta(t) : \mathcal{T} \to \mathbb{R}$:

$$\delta(t) = \sum_{k=1}^{K} \psi_k 1\{t = t_k\}.$$

Hence, we can rewrite the risk of estimator $\delta(T)$ when $T \sim N(b, 1)$ as

$$E_{T \sim N(b,1)} \left(\sum_{k=1}^{K} \psi_k 1\{T = t_k\} - b \right)^2. \tag{23}$$

Define $\mu_{kb} = \Pr_{T \sim N(b,1)}(T = t_k)$ as the probability of falling into the k'th grid point given bias b, which can be evaluated analytically via the following discrete approximation to the normal distribution

$$\mu_{kb} = \Phi\left((t_k + t_{k+1})/2 - b\right) - \Phi\left((t_k + t_{k-1})/2 - b\right), \tag{24}$$

where we define $t_0 = -\infty$ and $t_{K+1} = \infty$, which ensures that $\sum_{k=1}^{K} \mu_{kb} = 1$. The discretized approximation to the risk function (23) is therefore

$$\sum_{k=1}^{K} \psi_k^2 \mu_{kb} - 2b \sum_{k=1}^{K} \psi_k \mu_{kb} + b^2. \tag{25}$$

D.3 Computing minimax risk in the bounded normal mean problem

We now provide details on how to compute the minimax risk $r^{\mathrm{BNM}}(|\tilde{b}|)$ in the bounded normal mean problem, which allows us to easily compute the B-minimax risk as described in (5) for each $B \in \mathcal{B}$.

By definition, the minimax risk $r^{\text{BNM}}(|\tilde{b}|)$ is the minimized value of the following minimax problem

$$\min_{\delta} \max_{b \in [-|\tilde{b}|, |\tilde{b}|]} E_{T \sim N(b,1)} (\delta(T) - b)^2$$

whose solution is the minimax estimator $\delta^{\text{BNM}}\left(T; |\tilde{b}|\right)$. In particular, for each $|\tilde{b}| = B/\sigma_O \in \{0.1, 0.2, \ldots, 9\}$ we calculate the minimax risk $r^{\text{BNM}}(|\tilde{b}|)$ following the steps below. To compute the minimax risk function $r^{\text{BNM}}(|\tilde{b}|)$ for values of $|\tilde{b}|$ that are not included in the fine grid, we rely on spline interpolation.

1. Approximate the prior π with the finite dimensional vector $\pi \in \Delta^J$, where the parameter space $[-|\tilde{b}|, |\tilde{b}|]$ is approximated by an equally spaced grid of b values spanning $[-|\tilde{b}|, |\tilde{b}|]$ with a step size of 0.05, totaling to J grid values. Approximate the conditional risk function as in (25), where the support for $T \sim N(b, 1)$ is approximated by an equally spaced grid of t values spanning $[-|\tilde{b}| - 3, |\tilde{b}| + 3]$ with a step size of 0.1, totaling to K grid values. The minimax problem becomes

$$\max_{\pi \in \Delta^J} \min_{\{\psi_k\}_{k=1}^K} \sum_{\ell=1}^{J} \pi_\ell \left(\sum_{k=1}^{K} \psi_k^2 \mu_{kb_\ell} - 2b_\ell \sum_{k=1}^{K} \psi_k \mu_{kb_\ell} + b_\ell^2 \right). \tag{26}$$

2. The solution to the inner optimization yields the posterior mean $\psi_k^*(\pi) = \frac{\sum_{\ell=1}^{J} \pi_\ell \mu_{kb_\ell} b_\ell}{\sum_{\ell=1}^{J} \pi_\ell \mu_{kb_\ell}}$. The outer problem is then

$$\max_{\pi \in \Delta^J} \sum_{\ell=1}^{J} \pi_\ell \left(\sum_{k=1}^{K} \left(\psi_k^*(\pi)\right)^2 \mu_{kb_\ell} - 2b_\ell \sum_{k=1}^{K} \psi_k^*(\pi) \mu_{kb_\ell} + b_\ell^2 \right).$$

3. Solve the outer problem for the least favorable prior π^* based on sequential quadratic programming via MATLAB's `fmincon` routine. The minimax estimator $\delta^{\text{BNM}}\left(T; |\tilde{b}|\right)$ is therefore $\sum_{k=1}^{K} \psi_k^*(\pi^*) 1\{t = t_k\}$ and the minimax risk $r^{\text{BNM}}(|\tilde{b}|)$ is the minimized value.

Since the objective is concave in π (it is the pointwise infimum over a set of linear functions; see Boyd and Vandenberghe, 2004, p. 81), we can check that the algorithm has found a global maximum by checking for a local maximum.

D.4 Computing the optimally adaptive estimator for a given ρ^2

As explained in the main text, the adaptive problem only depends on Σ through the correlation coefficient ρ^2. For a given value of ρ^2, we use convex programming methods to solve for the function $\delta^*(t;\rho)$ based on the steps described below.

1. Approximate the prior π with the finite dimensional vector $\pi \in \Delta^J$, where the parameter space for b/σ_O is approximated by an equally spaced grid of $\tilde{b}$ values spanning $[-9, 9]$ with a step size of 0.025, totaling to J grid values. Approximate the conditional risk function as in (25), where the support for $T \sim N(\tilde{b}, 1)$ is approximated by an equally spaced grid of t values spanning $[-12, 12]$ with a step size of 0.05, totaling to K grid values. The adaptation problem (6) becomes

$$\max_{\pi\in\Delta^J} \min_{\{\psi_k\}_{k=1}^K} \sum_{\ell=1}^{J} \pi_\ell \omega_\ell \left(\sum_{k=1}^{K} \psi_k^2 \mu_{kb_\ell} - 2b_\ell \sum_{k=1}^{K} \psi_k \mu_{kb_\ell} + b_\ell^2 \right) + \rho^{-2} - 1 \tag{27}$$

   where $\omega_\ell = \left(r^{\mathrm{BNM}}(|\tilde{b}_\ell|) + \rho^{-2} - 1 \right)^{-1}$ using output from the previous subsection.

2. The solution to the inner optimization yields $\psi_k^*(\pi) = \frac{\sum_{\ell=1}^J \pi_\ell \mu_{kb_\ell} \omega_\ell b_\ell}{\sum_{\ell=1}^J \pi_\ell \mu_{kb_\ell} \omega_\ell}$. The outer problem is then

$$\max_{\pi\in\Delta^J} \sum_{\ell=1}^{J} \pi_\ell \omega_\ell \left(\sum_{k=1}^{K} \left(\psi_k^*(\pi)\right)^2 \mu_{kb_\ell} - 2b_\ell \sum_{k=1}^{K} \psi_k^*(\pi)\, \mu_{kb_\ell} + b_\ell^2 \right) + \rho^{-2} - 1.$$

3. Solve the outer problem for the least favorable (adaptive) prior π^* based on sequential quadratic programming via Matlab's fmincon routine. The adaptive estimator $\delta^*(t;\rho)$ is therefore $\sum_{k=1}^K \psi_k^*(\pi^*)\, 1\{t = t_k\}$. The loss of efficiency under adaptation is the minimized value.

As with the bounded normal mean problem, the objective is concave in π, so we can check that the algorithm has found a global maximum by checking for a local maximum.

This algorithm is a finite dimensional approximation to the optimization problem in Theorem 4.1(iii). While Theorem 4.1(iii) does not formally show the existence of a solution to this infinite dimensional problem, we find that the algorithm reliably converges to a global maximum, and that the least favorable prior stabilizes as the number of gridpoints

and range of the grid increase. Based on this numerical finding, we conjecture that the minimax problem in Theorem 4.1(iii) admits a least favorable prior, and that this solution can be approximated arbitrarily well using the our grid approach.

D.5 Computing the optimally adaptive estimator based on the lookup table

To simplify the computation of the optimally adaptive estimator, we pre-calculate the adaptive estimates over an unequally spaced grid $\tanh([0, 0.05, 0.10, \ldots, 3])$ of correlation coefficients using the algorithm described above. As ρ^2 approaches one, the solution becomes sensitive to small changes in ρ. The uneven spacing of the ρ grid allows for more accurate interpolation based on the simple pre-tabulated lookup table that we describe next.

To rapidly obtain a final estimator $\delta^*(T_O; \rho)$ for a given application, we conduct 2D interpolation across ρ^2 and t values to tailor the adaptive estimates to the exact parameter values desired. For example, we obtain $\delta^*(T_O; -0.524)$ based on spline interpolation at $\rho^2 = (-0.524)^2$ together with the observed test statistic T_O based on the 2D grid of ρ^2 and t values.

D.6 Computing the analytic adaptive estimators

To find the analytic adaptive estimators in the class of ERM estimators, soft thresholding estimators and hard thresholding estimators, it suffices to solve the two dimensional minimax problem in threshold λ and scaled bias level $\tilde{b}$. We provide details for the claim in the main text that this two dimensional minimax problem can be easily solved even though the minimax theorem does not apply to these restricted classes of estimators. To simplify the computation of the analytic adaptive estimator in practice, we pre-calculate the adaptive thresholds λ over an unequally spaced grid $\tanh([0, 0.05, 0.10, \ldots, 3])$ of correlation coefficients as explained above. To rapidly obtain a final estimator, for example, soft-thresholding estimator $\delta_{S,\lambda}(T_O; \rho)$ for a given application, we conduct a spline interpolation across ρ^2 values to tailor the threshold to the exact parameter values desired. For example, we obtain $\delta_{S,\lambda}(T_O; -0.524)$ firstly based on spline interpolation at $\rho^2 = (-0.524)^2$ to obtain the threshold λ, and then with the observed test statistic T_O.

The derivation for soft and hard thresholding is largely based on the following equality

using moments of a truncated standard normal $X_i \mid a < X_i < b$. Let $\phi(x)$ and $\Phi(x)$ denote the pdf and cdf of a standard normal distribution. Then for any $a < b$, we have

$$\int_a^b x^2\phi(x)dx = \Phi(b) - \Phi(a) - (b\phi(b) - a\phi(a)). \tag{28}$$

D.6.1 Soft thresholding

Rewrite the soft thresholding estimator as $\delta_{S,\lambda}(T_O) = \mathbf{1}\{T_O > \lambda\}(T_O - \lambda) + \mathbf{1}\{T_O < -\lambda\}(T_O + \lambda)$ and its risk function can be expressed as

$$\begin{aligned}
& E_{T_O\sim N(\tilde{b},1))}\left(\delta_{S,\lambda}(T_O) - \tilde{b}\right)^2 \\
= \; & E_{T_O\sim N(\tilde{b},1)}\left(\mathbf{1}\{T_O > \lambda\}\left(T_O - \lambda - \tilde{b}\right) + \mathbf{1}\{T_O < -\lambda\}\left(T_O + \lambda - \tilde{b}\right) - \mathbf{1}\{-\lambda < T_O < \lambda\}\tilde{b}\right)^2 \\
= \; & \tilde{b}^2\left(\Phi\left(\lambda - \tilde{b}\right) - \Phi\left(-\lambda - \tilde{b}\right)\right) + \int_{\lambda-\tilde{b}}^{\infty}(x-\lambda)^2\phi(x)dx + \int_{-\infty}^{-\lambda-\tilde{b}}(x+\lambda)^2\phi(x)dx
\end{aligned} \tag{29}$$

The integrals in (29) simplify to

$$\begin{aligned}
& \int_{\lambda-\tilde{b}}^{\infty}(x-\lambda)^2\phi(x)dx + \int_{-\infty}^{-\lambda-\tilde{b}}(x+\lambda)^2\phi(x)dx \\
= & \int_{\lambda-\tilde{b}}^{\infty}x^2\phi(x)dx + \int_{-\infty}^{-\lambda-\tilde{b}}x^2\phi(x)dx \\
& - 2\lambda\left(\int_{\lambda-\tilde{b}}^{\infty}x\phi(x)dx - \int_{-\infty}^{-\lambda-\tilde{b}}x\phi(x)dx\right) \\
& + \lambda^2\left(1 - \Phi\left(\lambda - \tilde{b}\right) + \Phi\left(-\lambda - \tilde{b}\right)\right) \\
= & 1 - \Phi\left(\lambda - \tilde{b}\right) + \Phi\left(-\lambda - \tilde{b}\right) + \left((\lambda-\tilde{b})\phi(\lambda-\tilde{b}) - (-\lambda-\tilde{b})\phi(-\lambda-\tilde{b})\right) \\
& - 2\lambda\left(\phi(\lambda-\tilde{b}) + \phi(-\lambda-\tilde{b})\right) + \lambda^2\left(1 - \Phi\left(\lambda - \tilde{b}\right) + \Phi\left(-\lambda - \tilde{b}\right)\right)
\end{aligned}$$

where we use the fact that $\int_{\lambda-\tilde{b}}^{\infty}x^2\phi(x)dx + \int_{-\infty}^{-\lambda-\tilde{b}}x^2\phi(x)dx = \int_{-\infty}^{\infty}x^2\phi(x)dx - \int_{-\lambda-\tilde{b}}^{\lambda-\tilde{b}}x^2\phi(x)dx$ and Equation (28).

The analytic adaptive objective function

$$\min_{\lambda}\max_{\tilde{b}}\frac{E_{T_O\sim N(\tilde{b},1))}\left(\delta_{S,\lambda}(T_O) - \tilde{b}\right)^2 + \rho^{-2} - 1}{r^{\mathrm{BNM}}(|\tilde{b}|) + \rho^{-2} - 1},$$

can now be easily solved by Matlab's `fminimax` function when the risk function is evaluated based on the simplified expression derived above, and the parameter space for $\tilde{b}$ is approximated by an equally spaced grid values spanning $[-9, 9]$ with a step size of 0.025.

D.6.2 Hard thresholding

Similarly rewrite hard thresholding as $\delta_{H,\lambda}(T_O) = (1 - \mathbf{1}\{-\lambda < T_O < \lambda\}) T_O$ and its risk function can be simplified due to Equation (28)

$$\begin{aligned} & E_{T_O \sim N(\tilde{b},1))} \left(\delta_{H,\lambda}(T_O) - \tilde{b}\right)^2 \\ = \; & E_{T_O \sim N(\tilde{b},1)} \left((1 - \mathbf{1}\{-\lambda < T_O < \lambda\}) \left(T_O - \tilde{b}\right) - \mathbf{1}\{-\lambda < T_O < \lambda\} \tilde{b}\right)^2 \\ = \; & \tilde{b}^2 \left(\Phi\left(\lambda - \tilde{b}\right) - \Phi\left(-\lambda - \tilde{b}\right)\right) + \int_{-\infty}^{\infty} x^2 \phi(x) dx - \int_{-\lambda-\tilde{b}}^{\lambda-\tilde{b}} x^2 \phi(x) dx. \end{aligned}$$

D.6.3 Adaptive ERM

For the adaptive ERM estimator $\delta_{ERM,\lambda}(T_O) = \frac{T_O^2}{T_O^2+\lambda} \cdot T_O$, we evaluate the risk function based on 10^5 simulations draws from $T_O \sim N(\tilde{b}, 1)$ and similarly optimize λ for the analytic adaptive objective function.

Appendix E Pooling controls (LaLonde, 1986)

LaLonde (1986) contrasted experimental estimates of the causal effects of job training derived from the National Supported Work (NSW) demonstration with econometric estimates derived from observational controls, concluding that the latter were highly sensitive to modeling choices. Subsequent work by Heckman and Hotz (1989) argued that proper use of specification tests would have guarded against large biases in LaLonde (1986)'s setting. An important limitation of the NSW experiment, however, is that its small sample size inhibits a precise assessment of the magnitude of selection bias associated with any given non-experimental estimator. In what follows, we explore the prospects of improving experimental estimates of the NSW's impact on earnings by utilizing additional non-experimental control groups and adapting to the biases their inclusion engenders.

We consider three analysis samples differentiated by the origin of the untreated ("control") observations. All three samples include the experimental NSW treatment group ob-

servations. In the first sample the untreated observations are given by the experimental NSW controls. In a second sample the controls come from LaLonde (1986)'s observational "CPS-1" sample, as reconstructed by Dehejia and Wahba (1999). In the third sample, the controls are a propensity score screened subsample of CPS-1. To estimate treatment effects in the samples with observational controls, we follow Angrist and Pischke (2009) in fitting linear models for 1978 earnings to a treatment dummy, 1974 and 1975 earnings, a quadratic in age, years of schooling, a dummy for no degree, a race and ethnicity dummies, and a dummy for marriage status. The propensity score is generated by fitting a probit model of treatment status on the same covariates and dropping observations with predicted treatment probabilities outside of the interval $[0.1, 0.9]$.

Let Y_U be the mean treatment / control contrast in the experimental NSW sample. We denote by Y_{R1} the estimated coefficient on the treatment dummy in the linear model described above when the controls are drawn from the CPS-1 sample. Finally, Y_{R2} gives the corresponding estimate obtained from the linear model when the controls come from the propensity score screened CPS-1 sample. We follow the applied literature in assuming trimming does not meaningfully change the estimand, a perspective that can be formalized by viewing the trimmed estimator as one realization of a sequence of estimators with trimming shares that decrease rapidly with the sample size (Huber et al., 2013).

Table A1 reports point estimates from all three estimation approaches along with standard errors derived from the pairs bootstrap. The realizations of (Y_{R1}, Y_{R2}) exactly reproduce those found in the last row of Table 3.3.3 of Angrist and Pischke (2009) but the reported standard errors are somewhat larger due to our use of the bootstrap, which accounts both for heteroscedasticity and uncertainty in the propensity score screening procedure. The realization of Y_U matches the point estimate reported in the first row of Angrist and Pischke (2009)'s Table 3.3.3 but again exhibits a modestly larger standard error reflecting heteroscedasticity with respect to treatment status.

While the experimental mean contrast (Y_U) of \$1,794 is statistically distinguishable from zero at the 5% level, considerable uncertainty remains about the magnitude of the average treatment effect of the NSW program on earnings. The propensity trimmed CPS-1 estimate lies closer to the experimental estimate than does the estimate from the untrimmed CPS-1 sample. However, the untrimmed estimate has a much smaller standard error than its trimmed analogue. Though the two restricted estimators are both derived from the CPS-1

Table A1: Estimates of the impact of NSW job training on earnings.

	Y_U	Y_{R1}	Y_{R2}	GMM_2	GMM_3	Adaptive	Pre-test
Estimate	1794	794	1362	1629	1210	1597	1629
Std error	(668)	(618)	(741)	(619)	(595)		
Max Regret	26%	∞	∞	∞	∞	7.77%	47.5%
Risk rel. to Y_U							
when $b_1 = 0$ and $b_2 = 0$	1	0.853	1.23	0.858	0.793	0.855	0.80
when $b_1 \neq 0$ and $b_2 = 0$	1	∞	1.23	0.858	∞	0.925	0.993
when $b_1 \neq 0$ and $b_2 \neq 0$	1	∞	∞	∞	∞	1.077	1.475

Notes: Bootstrap standard errors in parentheses computed using 1,000 bootstrap samples. The GMM_2 estimate imposes $b_2 = 0$ only while the GMM_3 estimate imposes $b_1 = 0$ and $b_2 = 0$. A J-test of the null $b_1 = b_2 = 0$ motivating GMM_3 yields a p-value at 0.04. A corresponding test of the null $b_2 = 0$ motivating GMM_2 yields a p-value of 0.51. "Risk rel. to Y_U" gives worst case risk scaled by the risk (i.e. variance) of Y_U. "Max regret" refers to the worst case adaptation regret in percentage terms $(A_{\max}(\mathcal{B}, \hat{\theta}) - 1) \times 100$.

sample, our bootstrap estimate of the correlation between them is only 0.75, revealing that each measure contains substantial independent information.

Combining the three estimators together via GMM, a procedure we denote GMM_3, yields roughly an 11% reduction in standard errors relative to relying on Y_U alone. However, the J-test associated with the GMM_3 procedure rejects the null hypothesis that the three estimators share the same probability limit at the 5% level ($p = 0.04$). Combining only Y_U and Y_{R2} by GMM, a procedure we denote GMM_2, yields a standard error 7% below that of Y_U alone. The J-test associated with GMM_2 fails to reject the restriction that Y_U and Y_{R2} share a common probability limit ($p = 0.51$). Hence, sequential pre-testing selects GMM_2.

Letting $b_1 \equiv \mathbb{E}[Y_{R1} - \theta]$ and $b_2 \equiv \mathbb{E}[Y_{R2} - \theta]$ our pre-tests reject the null that $b_1 = b_2 = 0$ and fail to reject that $b_2 = 0$. However, it seems plausible that both restricted estimators suffer from some degree of bias. The adaptive estimator seeks to determine the magnitude of those biases and make the best possible use of the observational estimates. In adapting to misspecification, we operate under the assumption that $|b_1| \geq |b_2|$, which is in keeping with the common motivation of propensity score trimming as a tool for bias reduction (e.g., Angrist and Pischke, 2009, Section 3.3.3). Denoting the bounds on $(|b_1|, |b_2|)$ by (B_1, B_2), we adapt over the finite collection of bounds $\mathcal{B} = \{(0, 0), (\infty, 0), (\infty, \infty)\}$, the granular nature of which dramatically reduces the computational complexity of finding the optimally adaptive estimator. Note that the scenario $(B_1, B_2) = (0, \infty)$ has been ruled out by assumption, reflecting the belief that propensity score trimming reduces bias. See Appendix F for further details.

From Table A1, the multivariate adaptive estimator yields an estimated training effect of \$1,597: roughly two thirds of the way towards Y_U from the efficient GMM_3 estimate. Hence, the observational evidence, while potentially quite biased, leads to a non-trivial (11%) adjustment of our best estimate of the effect of NSW training away from the experimental benchmark. In Table A2 we show that pairwise adaptation using only Y_U and Y_{R1} or only Y_U and Y_{R2} yields estimates much closer to Y_U. A kindred approach, which avoids completely discarding the information in either restricted estimator, is to combine Y_{R1} and Y_{R2} together via optimally weighted GMM and then adapt between Y_U and the composite GMM estimate. As shown in Table A3, this two step approach yields an estimate of \$1,624, extremely close to the multivariate adaptive estimate of \$1,597, but comes with substantially elevated worst case adaptation regret relative to a multivariate oracle who knows which pair of bounds in $\mathcal{B}$ prevails.

While the multivariate adaptive estimate of \$1,597 turns out to be very close to the pre-test estimate of \$1,629, the adaptive estimator's worst case adaptation regret of 7.7% is substantially lower than that of the pre-test estimator, which exhibits a maximal regret of 47.5%. The adaptive estimator achieves this advantage by equalizing the maximal adaptation regret across the three bias scenarios $\{(b_1 = 0, b_2 = 0), (b_1 \neq 0, b_2 = 0), (b_1 \neq 0, b_2 \neq 0)\}$ allowed by our specification of $\mathcal{B}$. When both restricted estimators are unbiased, the adaptive estimator yields a 14.5% reduction in worst case risk relative to Y_U. However, an oracle that knows both restricted estimators are unbiased would choose to employ GMM_3, implying maximal adaptation regret of $0.855/0.793 \approx 1.077$. When Y_{R1} is biased, but Y_{R2} is not, the adaptive estimator yields a 7.5% reduction in worst case risk. An oracle that knows only Y_{R1} is biased will rely on GMM_2, which yields worst case scaled risk of 0.858; hence, the worst case adaptation regret of not having employed GMM_2 in this scenario is $0.925/0.858 \approx 1.077$. Finally, when both restricted estimators are biased, the adaptive estimator can exhibit up to a 7.7% increase in risk relative to Y_U.

The near oracle performance of the optimally adaptive estimator in this setting suggests it should prove attractive to researchers with a wide range of priors regarding the degree of selection bias present in the CPS-1 samples. Both the skeptic that believes the restricted estimators may be immensely biased and the optimist who believes the restricted estimators are exactly unbiased should face at most a 7.7% increase in maximal risk from using the adaptive estimator. In contrast, an optimist could very well object to a proposal to rely on

Y_U alone, as doing so would raise risk by 26% over employing GMM_3.

Appendix F Details of bivariate adaptation

In Appendix E, we report the results of adapting simultaneously to the bias in two restricted estimators when the bias spaces take a nested structure. Denoting the bounds on $(|b_1|, |b_2|)$ of the two restricted estimators by (B_1, B_2), we adapt over the finite collection of bounds $\mathcal{B} = \{(0,0), (\infty, 0), (\infty, \infty)\}$. Note that the scenario $(B_1, B_2) = (0, \infty)$ has been ruled out by assumption, reflecting the belief that propensity score trimming reduces bias. The minimax risk over each bias space $\mathcal{C}_{(B_1,B_2)}$ is therefore

$$R^*(\mathcal{C}_{(B_1,B_2)}) = \begin{cases} \Sigma_U & \text{for } (B_1, B_2) = (\infty, \infty) \\ \Sigma_U - \Sigma_{UO,2}\Sigma_{O,2}^{-1}\Sigma_{UO,2} & \text{for } (B_1, B_2) = (\infty, 0) \\ \Sigma_U - \Sigma_{UO}\Sigma_O^{-1}\Sigma_{UO} & \text{for } (B_1, B_2) = (0, 0) \end{cases} \tag{30}$$

Then $\delta(Y_O)$ is the solution to the following problem

$$\inf_{\delta} \max_{(B_1,B_2)\in\mathcal{B}} \frac{\max_{b\in\mathcal{C}_{(B_1,B_2)}} E_{Y_O\sim N(b,\Sigma_O)}(\delta(Y_O) - \Sigma_{UO}\Sigma_O^{-1}b)^2 + \Sigma_U - \Sigma_{UO}\Sigma_O^{-1}\Sigma_{UO}}{R^*(\mathcal{C}_{(B_1,B_2)})}$$

Since the three spaces are nested, we can rewrite the adaptation problem as

$$\inf_{\delta} \sup_{b\in\mathbb{R}\times\mathbb{R}} \frac{E_{Y_O\sim N(b,\Sigma_O)}(\delta(Y_O) - \Sigma_{UO}\Sigma_O^{-1}b)^2 + \Sigma_U - \Sigma_{UO}\Sigma_O^{-1}\Sigma_{UO}}{\tilde{R}(\tilde{\mathcal{S}}(b))}$$

where the scaling is

$$\tilde{R}(\tilde{\mathcal{S}}(b)) = \begin{cases} \Sigma_U - \Sigma_{UO}\Sigma_O^{-1}\Sigma_{UO} & \text{if } b_1 = b_2 = 0 \\ \Sigma_U - \Sigma_{UO,2}\Sigma_{O,2}^{-1}\Sigma_{UO,2} & \text{if } b_1 \neq 0, b_2 = 0 \\ \Sigma_U & \text{if } b_1 \neq 0, b_2 \neq 0 \end{cases} \tag{31}$$

Given the high dimensionality of the adaptation problem, we use CVX instead of Matlab's fmincon to solve the scaled minimax problem.

F.1 Pairwise adaptation

For comparison with the trivariate adaptation estimates reported in the text, we also consider pairwise adaptation using only Y_U and Y_{R1} or only Y_U and Y_{R2}, keeping the bias spaces as before. Specifically to adapt using only Y_U and Y_{Rj}, we consider an oracle where the set $\mathcal{B}$ of bounds B on the bias consists of the two elements 0 and ∞.

Table A2: Pairwise adaptive estimates

	Y_U	Y_R	GMM	Adaptive	Soft-threshold	Pre-test
CPS-1 untrimmed	1794	794	1123	1659	1608	1794
Std error	(668)	(617)	(600)			
Rel. risk when $b = 0$	1	0.85	0.81	0.863	0.869	0.894
Rel. risk when $b \neq 0$	1	∞	∞	1.071	1.078	1.541
Max Regret	24%	∞	∞	7.1%	7.8%	54%
Max Regret (rel. to multivariate)	26%	∞	∞	24.8%	25.6%	79.5%
Threshold					0.63	1.96
CPS-1 trimmed	1794	1362	1629	1657	1638	1362
Std error	(668)	(741)	(619)			
Rel. risk when $b = 0$	1	1.23	0.86	0.9	0.91	1.166
Rel. risk when $b \neq 0$	1	∞	∞	1.05	1.055	2.051
Max Regret	16.4%	∞	∞	5%	5.5%	105%
Max Regret (rel. to multivariate)	26%	∞	∞	13.6%	14.2%	105%
Threshold					0.62	1.96

Notes: Bootstrap standard errors in parentheses computed using 1,000 bootstrap samples. In the top panel Y_R corresponds to estimates using the untrimmed CPS-1 as controls, which are referred to as Y_{R1} in the main text. In the bottom panel, Y_R corresponds to estimates derived from the propensity score trimmed CPS-1 sample, which are referred to as Y_{R2} in the main text. Adaptive estimates adapt pairwise between Y_U and Y_R within panel. If applicable, the adaptive thresholds are reported. "Max regret" refers to the worst case adaptation regret in percentage terms $(A_{\max}(\mathcal{B}, \hat{\theta}) - 1) \times 100$. "Max Regret (rel. to multivariate)" refers to the worst case adaptation regret in terms of the multivariate oracle. "Rel. risk" gives worst case risk scaled by the risk (i.e. variance) of Y_U. The correlation between Y_U and $Y_{Rj} - Y_U$ is -0.44 in the top panel and -0.38 in the bottom panel.

Table A2 shows that pairwise adaptation produces estimates much closer to Y_U than the multivariate adaptive estimate. While pairwise adaptive estimates both incur smaller adaptation regret, the efficiency gain when the model is correct is smaller than with the multivariate adaptive estimate.

Table A3: Adapting pairwise with GMM composite

	Y_U	Y_{comp}	GMM	Adaptive	Soft-threshold	Pre-test
Estimate	1794	882	1173	1624	1601	1794
Std error	(668)	(612)	(595)			
Max Regret	26%	∞	∞	8%	8.3%	56%
Max Regret (rel. to multivariate)	26%	∞	∞	25.4%	26.3%	81.5%
Threshold			∞		0.64	1.96

Notes: Adaptive estimates for the impact of job training, adapting to $B_{\text{comp}} \in \{0, \infty\}$, which is the bound on the bias of the composite estimator $Y_{\text{comp}} = \arg\min_\theta (Y_R - \theta)' \Sigma_R^{-1} (Y_R - \theta)$. GMM combines Y_{comp} and Y_U optimally under the assumption that Y_{comp} is unbiased. If applicable, the adaptive thresholds are reported. "Max regret" refers to the worst case adaptation regret in percentage terms $(A_{\max}(\mathcal{B}, \hat{\theta}) - 1) \times 100$. "Max Regret (rel. to multivariate)" refers to the worst case adaptation regret relative to the multivariate oracle in (30). The correlation coefficient between Y_U and $Y_{\text{comp}} - Y_U$ is -0.45.

F.2 Bivariate adaptation with GMM composite

For another comparison with the trivariate adaptation estimates reported in the text, we also consider combining Y_{R1} and Y_{R2} first via optimally weighted GMM, which is a composite of the two Y_{comp}. We then adapt between Y_U and Y_{comp}. The bias space is now also a composite of the two-dimensional bias space $\mathcal{C}_{(B_1,B_2)}$, and we consider an oracle where the set $\mathcal{B}$ of bounds B on the bias consists of the two elements 0 and ∞.

Table A3 shows that composite adaptation produces estimates very similar to the multivariate adaptive estimate. The adaptation regret relative to an oracle who knows a bound on the bias of composite is also small. However, for a fair comparison with multivariate adaptation, one should compare its efficiency loss relative to the multivariate oracle with minimax risk specified in (30). This notion of worst case regret is substantially higher at 25% because bivariate adaptation against the GMM composite cannot leverage the nested structure of the multivariate parameter space $\mathcal{B}$.

References for Online Appendix

Angrist, J. D. and J.-S. Pischke (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Bickel, P. J. (1983). Minimax estimation of the mean of a normal distribution subject to

doing well at a point. In M. H. Rizvi, J. S. Rustagi, and D. Siegmund (Eds.), *Recent Advances in Statistics*, pp. 511–528. Academic Press.

Boyd, S. P. and L. Vandenberghe (2004, March). *Convex Optimization*. Cambridge University Press.

Dehejia, R. H. and S. Wahba (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association 94*(448), 1053–1062.

Heckman, J. J. and V. J. Hotz (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American statistical Association 84*(408), 862–874.

Huber, M., M. Lechner, and C. Wunsch (2013). The performance of estimators based on the propensity score. *Journal of Econometrics 175*(1), 1–21.

Johnstone, I. M. (2019). *Gaussian estimation: Sequence and wavelet models*. Online manuscript available at https://imjohnstone.su.domains/.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, 604–620.

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley & Sons.